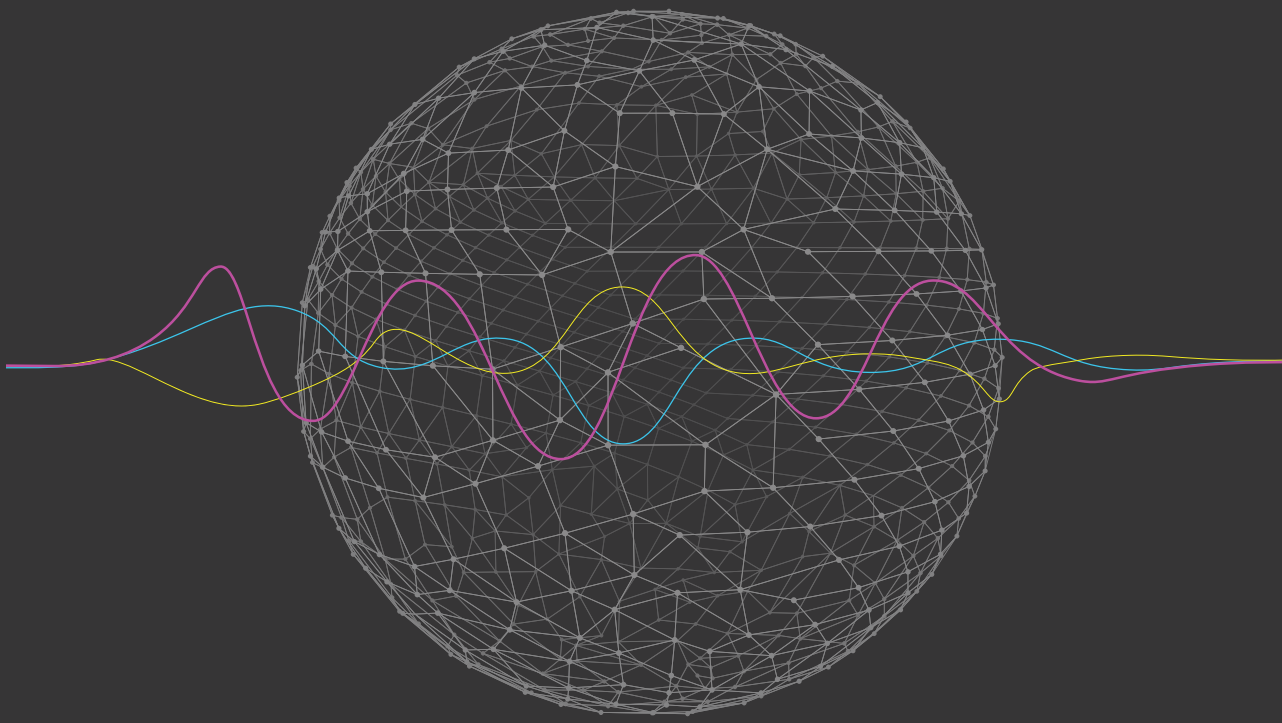


From chromatin to gene regulatory networks in embryonic development and evolution

Georgios Georgiou

From chromatin to gene regulatory networks in embryonic development and evolution

Georgios Georgiou

Colofon

From chromatin to gene regulatory networks in embryonic development and evolution,
Georgios Georgiou
ISBN: 978-94-028-1646-4

Cover design by: Mr. Panda Design Studio www.mrpandadesignstudio.com
Layout and design by Anna Bleeker | persoonlijkproefschrift.nl
Printed by Ipskamp Printing, proefschriften.net

From chromatin to gene regulatory networks in embryonic development and evolution

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 23 september 2019
om 14.30 uur precies

door
Georgios Georgiou
geboren op 31 augustus 1986
te Nicosia, Cyprus

Promotor:
Prof.dr. G.J.C. Veenstra

Copromotor:
Dr. S.J van Heeringen

Manuscriptcommissie:
Prof. dr. B. Franke
Prof. dr. A.H. van Kampen (Universiteit van Amsterdam)
Dr. C.F.H.A. Gilissen

TABLE OF CONTENTS

CHAPTER ONE

Introduction

For every multicellular animal, life starts as a single cell. The development of a single cell into a complex and multicellular organism is a fascinating process of nature. A single fertilized egg will undergo many cell divisions, differentiating eventually into a diversity of complex cell types, organized into organs and tissues. While nearly every cell type contains a copy of the exact same genetic information, they have a unique transcriptional program which leads to diverse phenotypes. The genetic information is stored as deoxyribonucleic acid (DNA); a double helix structure with two strands complementary to each other. In the nucleus of eukaryotic cells, DNA is compacted with histone proteins in a structure termed chromatin.

The processes of development and differentiation are made possible by precise regulation of gene expression through promoters and enhancers. Promoters are DNA sequences proximal to the transcription start sites (TSS) of genes, and they promote transcription. The core promoter directs the initiation of transcription by serving as a docking site for all the necessary components of the transcription machinery. The core promoter comprises several DNA sequences, such as the TATA box and CpG islands, which can be bound by co-factors and the general transcription factors (Matsui et al. 1980; Struhl 1995). These factors, together with RNA Polymerase II (RNAPII), form the preinitiation complex (PIC) which directs the RNAPII to the nearby transcription start site (Lee and Young 2000; Roger D. Kornberg 2007; Kim et al. 1997; Murakami et al. 2015). RNAPII is recruited by the initiation factors to the core promoter region of genes and subsequently initiates, elongates and terminates transcription (Matsui et al. 1980; Zawel and Reinberg 1993). Another element of the PIC is the multi-subunit mediator complex. The mediator is an evolutionarily conserved protein complex, with almost 30 polypeptides in humans and 25 in yeast (Poss et al. 2013; Yin & Wang 2014; Soutourina 2018). It is required for transcription and functions as a 'bridge' between transcription factors and basal transcriptional machinery. The mediator complex is recruited to enhancers via direct interactions with transcription factors bound in those regions and through chromatin looping interacts with PIC.

Transcription factors are proteins capable of affecting gene expression. Sequence-specific transcription factors bind to promoters or to distal regulatory regions, called enhancers. Enhancers are usually located up to 1 megabase pairs (Mb) away from promoters and transcription start sites (Pennacchio et al. 2013). In the nucleus, the DNA is packaged in chromatin fibers. This three-dimensional DNA structure facilitates the physical interaction between enhancers and promoters. Complex gene regulatory networks (GRNs), encoded in the genome, regulate the precise spatial and temporal control of gene expression. Transcription factors interact with chromatin and their target genes, orchestrating the development.

This chapter aims to introduce the main concepts relating to chromatin regulation and GRNs during the development of western clawed frog (*Xenopus tropicalis*) and African clawed

frog (*Xenopus laevis*). These processes are described individually, although they are highly interconnected.

1. EMBRYONIC DEVELOPMENT IN XENOPUS

During the early stages of embryonic development, due to common ancestry, vertebrates go through broadly similar developmental processes (Wolpert, Tickle, and Arias 2015; Gilbert 2000). This process is called embryogenesis and starts with the fertilization of an egg by a sperm cell, followed by cleavage, gastrulation and organogenesis. Mammals, such as humans and mice, develop the fertilized eggs inside the female's body, while other vertebrates, like amphibians, fishes and reptiles, lay eggs in water or on land. Amphibians are often used to study embryonic development, because of their external fertilization and the large number of eggs. Two commonly used models for vertebrate development are the western clawed frog *Xenopus tropicalis* and the African clawed toad *Xenopus laevis*, as they have several anatomical, physiological and genetic similarities with humans (Wheeler and Brändli 2009; Schmitt, Gull, and Brändli 2014; Hempel and Kühl 2016).

In *Xenopus*, the process of embryonic development starts from a fertilized egg with cleavage, a series of rapid cell divisions. During cleavage, cells start synchronously dividing into smaller cells, referred to as blastomeres. The embryo forms into a sphere and inside develops the blastocoel, a fluid-filled cavity. At this stage, the embryo becomes what is known as the blastula. After the 12 initial divisions, at the mid-blastula stage, the embryo has three regions; the animal cap, the equatorial or marginal zone and the vegetal mass. At the end of the blastula stage, the embryo reorganizes into the three germ layers; ectoderm, endoderm, and mesoderm. Ectoderm is the layer which forms tissues and organs on the outside; thus the name ektos (outside) and derma (skin). Some of the organs and tissues derived from the ectoderm are skin, nervous system, cornea, lens and epithelial lining of the mouth. Endoderm comes from the Greek words entos (inside) and derma and is the layer which moves on the inside of the embryo. Out of it derives the epithelial lining of the digestive tract, respiratory system, liver, pancreas, and reproductive system. Mesoderm, mesi (middle) and derma, among others, is also responsible for the formation of muscles, septa, skeleton, mesenteries, and reproductive system. The stage during which the three germ layers can be distinguished for the first time is known as the gastrula.

In the gastrula-stage embryo, the involution of endodermal and mesodermal cells forms a pit-like region called the blastopore. The blastopore region is vital for the embryo because the dorsal blastopore lip is the region of the Spemann organizer (Spemann and Mangold 1924). The Spemann organizer is a cluster of cells which induce the dorsal-ventral axis and neural tissues

(Crease, Dyson, and Gurdon 1998). Following the gastrulation stage comes the neurulation stage where the mesoderm gives rise to the notochord. The notochord directs the formation of the neural tube from ectodermal cells by folding the neural plate. The neural plate folds inwards, forming two parallel folds which eventually merge and form the neural tube, which separates from ectoderm. The neural tube will ultimately form the spinal cord and brain.

During the tailbud stage, following neurulation, organogenesis starts with the formation of the tail and leads to the development of other organs and tissues. Organs are formed from cells derived from the three germ layers. The ectoderm will give rise to the epidermis and nervous system, the endoderm to gastrointestinal, respiratory and urinary systems and the mesoderm to the notochord, cartilage, connective tissue, trunk muscles, kidneys and blood (Blitz, Andelfinger, and Horb 2006; Kiecker, Bates, and Bell 2016). After organs are formed, the metamorphosis stage starts, where the embryo develops into a tadpole and eventually transforms into the adult form.

The different cell types that arise during development are made possible by the precise control of gene expression. Protein complexes interact with each other and with the DNA and the chromatin and modulate the transcription of genes. The next section will focus on the mechanisms of regulation.

2. CHROMATIN AND REGULATION OF TRANSCRIPTION

2.1 Chromatin and the epigenetic code

The genetic information of every metazoan organism is stored in its DNA (Hershey and Chase 1952; Avery, Macleod, and McCarty 1944). Every cell shares a copy of approximately the same information, stored in the nucleus. In humans, the total length of DNA molecules in a single cell is more than two meters. To fit into the nucleus, the DNA is wrapped around histones in a spool-like unit called the nucleosome (R. D. Kornberg and Thomas 1974). Nucleosomes consist of two copies of four histone proteins (H2A, H2B, H3, and H4), with 145-147 base pairs of DNA wrapped around them (Richmond et al. 1984; Luger et al. 1997). The resulting octameric complex of histone proteins and DNA is known as chromatin (Figure 1) (Flemming 1882). Based on the compaction of the DNA with nucleosomes, chromatin can be classified either as heterochromatin or euchromatin. In heterochromatin, chromatin is compact and inaccessible for transcription. Heterochromatin is found in sequences areas that are highly condensed and rich in repetitive sequences, such as centromeres and telomeres, and it is often associated with repression of transcription (Nishibuchi and Déjardin 2017). Conversely, euchromatin has relatively loose compaction. It is enriched with actively transcribed genes and is often associated with activation of transcription (Kouzarides 2007).

Apart from acting as packing scaffolds, nucleosomes can influence the activation and repression of gene transcription. Residues in histones and histone tails can be subject to covalent post-translational modifications (PTMs) (Allfrey, Faulkner, and Mirsky 1964). Histone amino N-terminal tails interact with their adjacent nucleosomes in over 60 residues (Kouzarides 2007). Modifications include acetylation and methylation of lysines and arginines, phosphorylation of serines, threonines and tyrosines and ubiquitylation and sumoylation of lysines (Chrun, Modolo, and Daniel 2017; Bannister and Kouzarides 2011). These modifications can affect gene expression, DNA repair, replication, and recombination by recruiting remodeling enzymes and altering chromatin structure (Bannister and Kouzarides 2011). These histone enzymes can be distributed into three main categories: writers, readers and erasers. Writers, such as histone acetyltransferases (HATs) and histone methyltransferases (HMTs), are enzymes that add PTMs to histones (Gillette and Hill 2015). On the other hand, erasers, such as histone deacetylases (HDACs) and histone demethylases, are enzymes which remove PTMs from histones (Gillette and Hill 2015). Last, readers, such as bromo and chromo domains, are acetyl- or methyl-binding proteins that recognize specific or combination of PTMs on histones and govern transcription (Gillette and Hill 2015; Xu et al. 2017).

The introduction of high-throughput sequencing allowed the genome-wide mapping of histone-associated PTMs and their correlation with transcriptional activity. Acetylation has been shown to be associated with transcriptional activation, whereas methylation is associated with both activation and repression (Kouzarides 2007). Monomethylation of histone H3 at lysine 4 (H3K4me1), trimethylation of histone H3 protein at of the lysine 4 (H3K4me3) and trimethylation of histone H3 at lysine 36 (H3K36me3) are associated with transcriptional activation. H3K4me1 is found in permissive enhancers, whereas H3K4me3 is located at active promoters and H3K36me3 is associated with gene bodies of actively transcribed genes. Dimethylation of histone H3 at lysine 9 (H3K9me2), trimethylation of histone H3 at lysine 9 (H3K9me3) and trimethylation of histone H3 at lysine 27 (H3K27me3) are associated with heterochromatin regions and transcriptional repression.

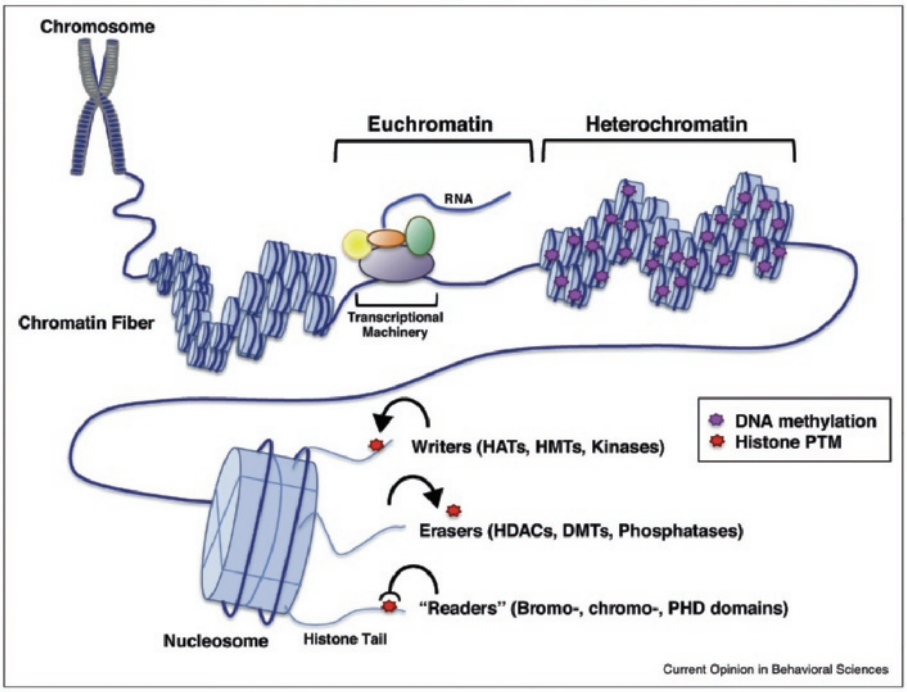


Figure 1. Epigenetic regulation of gene expression. In chromosomes, DNA is wrapped around nucleosomes in a structure called chromatin. Based on the compaction, chromatin either defined as euchromatin or heterochromatin. Writers, erasers and readers are histone enzymes that can respectively deposit, remove and read the PTMs. Reprinted from Current Opinion in Behavioral Sciences, 25, Ryan M Bastle, Ian S. Maze, Chromatin regulation in complex brain disorders, pp. 57-65, Copyright (2019), with permission from Elsevier.

Chromatin immunoprecipitation with modification-specific antibodies followed by sequencing (ChIP-seq) can generate high-resolution, genome-wide histone modification profiles (Nakato and Shirahige 2017). Histone proteins are crosslinked to DNA and using histone-specific antibodies ChIP-seq provides a snapshot of the interactions happening in the genome at that specific time point. Combinations of histone modifications, or chromatin states, have been associated with regulatory elements, such as enhancers, promoters, transcribed or repressed regions and other novel classes of elements (Ernst and Kellis 2017). Identification of those states can lead to a better annotation of coding and non-coding genomic regions and help with the understanding of gene regulation and cellular differentiation and genome-wide association studies.

2.2 Transcription factors

The process of gene regulation controls the development of a multicellular organism from a single fertilized egg. Gene regulation is governed by an interplay between chromatin and transcription factors (TFs). The term "transcription factors" is used to describe proteins that affect gene expression by activating or repressing transcription (Spitz and Furlong 2012). In the human genome, about 8% of the genes are considered to encode TFs, with some expressed only in specific cell types or at specific developmental stages (Lambert et al. 2018). In prokaryotes, a single TF can drive a program of gene expression, however, in eukaryotes multiple TFs tend to form complex GRNs that guide gene expression (Levine and Tjian 2003).

Sequence-specific TFs are DNA-binding proteins that bind to DNA by recognizing specific DNA sequences, also known as DNA motifs, and function as activators or repressors and recruit other co-activators or co-repressors. The size of motifs varies from 4 up to ~30 bp, however, in eukaryotes motifs tend to be between 6 and 12 bp long (Spitz and Furlong 2012; Tuğrul et al. 2015; A. Khan et al. 2018; Weirauch et al. 2014). While the exact function of TFs might be different, the binding domains are usually highly conserved among species (Borneman et al. 2007; Bradley et al. 2010; Schmidt et al. 2010; He et al. 2011). TFs can bind in promoter regions close to the transcription start sites of genes they regulate or in distal locations, in cis-regulatory elements thousands of bp away acting as enhancers or silencers (Heintzman et al. 2009; Stender et al. 2010; Heinz et al. 2010). The majority of TFs bind in open chromatin, however, some TFs, known as pioneer factors, are capable of binding in compact chromatin (Zaret and Carroll 2011; Iwafuchi-Doi and Zaret 2014; Drouin 2016). Pioneer TFs recruit other TFs and chromatin remodeling enzymes, resulting in opening the chromatin which is associated with DNA regulatory elements, including active promoters and enhancers (Lupien et al. 2008; Iwafuchi-Doi and Zaret 2014; Drouin 2016; Gross and Garrard 1988; Cockerill 2011).

2.3 Enhancers

Enhancers are cis-regulatory regions, usually located up to 1 megabase pairs (Mb) away from promoters and transcription start sites (Pennacchio et al. 2013). Enhancers got their name because of their ability to enhance gene transcription (Banerji, Rusconi, and Schaffner 1981). Their size varies from 50 bp to 1.5 kilobase pairs (kb) and they are packed with transcription factor binding sites (Blackwood and Kadonaga 1998; Pennacchio et al. 2013; ENCODE Project Consortium 2012). Without regard to orientation and over long distances, enhancers come in contact with the promoters through a DNA looping mechanism and ultimately enhance gene transcription (Banerji, Rusconi, and Schaffner 1981; Maston, Evans, and Green 2006). Long-range interactions of enhancers are often restricted by topologically associated domains (TADs) (Dixon et al. 2012). TADs are continuous genomic regions, defined by insulator elements, where regulatory elements, like enhancers and promoters, interact within (Pombo and Dillon 2015; de

Laat and Duboule 2013; Dixon et al. 2012). A single enhancer can regulate multiple genes and multiple enhancers can regulate a single gene (Mohrs et al. 2001; Li et al. 2012; Pennacchio et al. 2013). The number of enhancers regulating a gene often depends on the function of the gene. Tissue-specific genes often have multiple enhancers, whereas some ubiquitously expressed genes have no enhancers (Zabidi et al. 2015).

Enhancers can be identified using a range of methods, based on their properties, such as chromatin state and sequence composition. Recent advancements in DNA sequencing and computational methods have facilitated the identification of enhancers on a genome-wide level. The histone modifications H3K4me1 and H3K27ac are found to be associated with putative enhancers (Heintzman et al. 2007; Rada-Iglesias et al. 2011; Bonn et al. 2012). Using ChIP-seq to build genome-wide histone modification profiles and software such as Segway and ChromHMM, the presence of H3K4me1 and H3K27ac can define enhancers states (Hoffman et al. 2012; Ernst and Kellis 2012). Based on patterns of those histone modification marks, enhancers can be assigned in four categories; poised, primed, active and latent.

Poised enhancers are enhancers which are characterized by reduced chromatin accessibility. They are found to overlap with H3K4me1 and the polycomb-associated mark H3K27me3 (Creyghton et al. 2010; Shlyueva, Stampfel, and Stark 2014). They are located in euchromatic regions and are poised for activation in later developmental stages (Shlyueva, Stampfel, and Stark 2014). Primed enhancers can also be located in nucleosome-free regions, lacking H3K27me3, and are poised for activity in later stages of the development. Recent studies on mouse embryonic stem cells (mESCs) show that activation of primed enhancers depends on the UTX (H3K27 demethylase)-MLL4 (H3K4 methyltransferase) complex and the histone acetyltransferase p300 (S.-P. Wang et al. 2017). The UTX acts as a recruitment mechanism for MLL4 and p300 facilitates the deposition of MLL4-depended H3K3me1, which subsequently boosts the deposition of p300-mediated H3K27ac. Active enhancers are typically located in DNA accessible regions, depleted of nucleosomes, and bound by TFs and the coactivator p300. They are found to overlap with H3K4me1 and H3K27ac (Creyghton et al. 2010; Shlyueva, Stampfel, and Stark 2014). Binding of TFs generally determines the activity of active enhancers (Calo and Wysocka 2013). Using techniques as DNase-seq (DNase I hypersensitive sites sequencing) and ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) to identify DNA accessible regions can lead to the identification of enhancers, although, there is not always an association between DNA accessible regions and active enhancers (Song and Crawford 2010; Buenrostro et al. 2015). Often DNA accessible regions can be enriched as well with promoters, silencers, and insulators (Chatterjee and Ahituv 2017; Shlyueva, Stampfel, and Stark 2014). Latent enhancers are located on compact chromatin and unmarked by any histone modifications. However, in response to external stimuli, such as a pioneer TF or induction of cellular signaling

pathways, the chromatin opens and facilitates the deposition of H3K4me1 and H3K27ac (Ostuni et al. 2013; Shlyueva, Stampfel, and Stark 2014).

Another way to identify enhancers is by the presence of the histone acetyltransferase p300 coactivator (Heintzman et al. 2007; Q. Wang, Carroll, and Brown 2005). P300 can catalyze the deposition of the active enhancer mark H3K27ac (Pasini et al. 2010; Jin et al. 2011). The gene *Ep300* encodes p300, which is recruited to enhancers by TFs (Eckner et al. 1994; Chan and La Thangue 2001). P300 is referred to as a transcriptional coactivator because of its function to bind to TFs and activate transcription.

The binding of TFs generally determines the activity of enhancers. Computational and experimental approaches to build a genome-wide landscape of TF binding sites (TFBSs) can contribute to the identification of active enhancers. Computational approaches rely on the identification of TFBSs (methods described in 1.3.2). Based on the premise that enhancers contain an abundance of TFBSs, these approaches aim to discover putative enhancers by looking for genomic regions enriched in TFBSs (Berman et al. 2002). Along the same line, TFBSs conserved among species are identified using methods based on multiple sequence alignment (Kheradpour et al. 2007; Del Bene et al. 2007).

The terms "redundant enhancers" or "shadow enhancers" were initially used to describe enhancers in *Drosophila* embryos which are located further away from their target genes and diverged twice as fast (Hong, Hendrix, and Levine 2008). They have similar transcription patterns compared to the primary enhancers and are bound by the same TFs. Loss of one of the enhancers does not have a significant effect on the gene expression nor affects the viability of the embryo (Perry et al. 2010). However, they cannot be considered redundant. Experiments have shown that they have an essential role in development by acting as a canalization mechanism and buffering environmental and genetic perturbations (Perry et al. 2010; Cannavò et al. 2016).

Another class of enhancers that believed to be redundant are the stretch enhancers or super-enhancers. The terms stretch enhancers or super-enhancers are used for describing enhancers found in clusters (Parker et al. 2013; Whyte et al. 2013). Super-enhancers are large enhancer domains, usually more than ten kilobases long, with high enhancer activity. Like typical enhancers, super-enhancers are located in DNA accessible regions and found to be enriched with H3K27ac, H3K4me1, coactivators and cell-type specific master regulators, such as Oct4, Sox2, and Nanog in embryonic stem cells (ESCs) (Whyte et al. 2013; Lovén et al. 2013). In contrast with conventional enhancers, super-enhancers have unusually high levels of the mediator protein MED1 (Lovén et al. 2013; Whyte et al. 2013). Moreover, super-enhancers in ESCs

are enriched with the binding of KLF4 and ESRRB, which have been shown to have a crucial role in pluripotency and reprogramming (Niederriter et al. 2015; Festuccia et al. 2012). They have been found near genes coding for developmental regulators and genes implicated in cell identity. However, despite these claims, the function of super-enhancers is still controversial. Recent studies have shown that super-enhancers are groups of regular enhancers, not functionally distinct from others, acting redundantly (Moorthy et al. 2017; Xie et al. 2017). It remains unclear if super-enhancers can be described as new distinct functional elements with their properties or just as a group of typical enhancers with stronger activity (Pott and Lieb 2015; Hay et al. 2016; Huang et al. 2018).

The term "sentences" is often used to describe enhancers, where the "words" consist of TFBSs. As in every language, the understanding of the formation of sentences is crucial to understand the "grammar" behind enhancers. Enhancer "grammar" describes the rules that govern their composition regarding TFBSs. The content, arrangement, orientation, and affinity of TFBSs affect gene transcription; therefore, breaking down enhancer "grammar" to "words" is an essential step towards understanding their function and explaining transcriptional regulation. The next section focuses on the existing methods of representing and identifying TFBSs.

3. TRANSCRIPTION FACTOR BINDING MOTIFS

3.1 Representation of motifs

There are several ways to represent a TFBS. Each one of them aims to represent a sequence pattern, which makes it distinguishable from the surrounding sequence. The most straightforward model is the consensus sequence (string representation) (S. Sinha and Tompa 2000; Saurabh Sinha 2003; Pesole et al. 1992; Buhler and Tompa 2002; H. C. M. Leung and Chin, n.d.). This model is rather simple, compact and similar to regular expressions. It describes the motifs of interest using a string of n nucleotides (Figure 2A). The consensus sequence representation shows the most frequent nucleotide at each position of the sequence(s) or IUPAC symbols to represent ambiguous nucleotides (Figure 2B) (Cornish-Bowden 1985).

Another widely-used approach to represent motifs is the frequency or likelihood matrix (Position Probability Matrix (PPM), Position Weight Matrix (PWM) or Position Specific Scoring Matrix (PSSM)) (Stormo et al. 1982; Henry C. M. Leung, Chin, and Chan 2007; Timothy L. Bailey and Elkan 1995). The PSSM consists of an N by M matrix, where M is the length of the sequence and N the number of possible nucleotides (A, C, G, and T). The probability of each nucleotide at each position is calculated by counting the frequency of each nucleotide at each position and normalizing it by the number of sequences (Figure 2C). To transform the probabilities to

weights, each probability is normalized by the frequency of each nucleotide in the background and then transformed using log2 likelihood. Because of its descriptive power, sensitivity, and precision, the PSSM became a more popular model to represent motifs. Matrices can be visually represented using sequence logos (Figure 2D). Sequence logos were created in 1990 by Thomas D. Schneider for displaying patterns in a set of aligned sequences (Schneider and Stephens 1990). Logos have the length of the corresponding sequence and at each position a stack of letters representing the four nucleotides. The size of each nucleotide is often displayed in bits and is an indicator of the frequency and information content, with the most frequent located at the top. Adaptations of sequence logos attempt to offer more insights from motif representation by considering more biophysical mechanisms, such as the actual relative free energy of binding, methylation sensitivity, phosphate linkage and DNA shape (Foat, Morozov, and Bussemaker 2006; Kribelbauer et al. 2017; Fortin, Schulze, and Babbitt 2015; Yang et al. 2017).

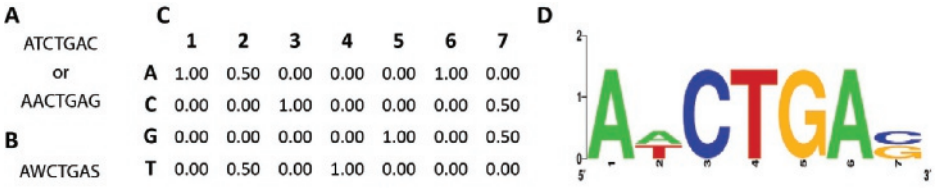


Figure 2. Examples of different motif representations. A) The preferred nucleotide sequences using a string of seven nucleotides. **B)** The consensus sequence using IUPAC symbols to represent ambiguous nucleotides. **C)** The Position Probability Matrix (PPM) representing the probability of each nucleotide for each position. D) The sequence logo of the PPM in C.

The PSSM model is based on the assumption that mononucleotides have an independent effect on binding affinity. However, this is a simplification and alternative models have been developed that take into consideration the dependencies between nucleotides and other parameters, such as the role of DNA shape and electrostatic potential or the impact of DNA methylation (Bulyk, Johnson, and Church 2002; Jolma et al. 2013; Rohs et al. 2009; Yin et al. 2017). However, due to the complexity and the sensitivity to parameter-specific noise, these additive models failed to become widely accepted. Increasing the variables taken into account increases the amount of data needed to estimate the parameters and, therefore, the computational time (Benos, Bulyk, and Stormo 2002). Furthermore, depending on the TFs used, models can have different performance (Weirauch et al. 2013).

In summary, models taking into account different parameters are not robust and come with an increase in complexity and computational cost; hence PWMs are still the preferred and most common model to represent TFBS (Benos, Bulyk, and Stormo 2002; Weirauch et al. 2013). The next section will focus on the approaches used to identify TFBS.

3.2 Identification of transcription factor binding motifs

TFBSs can be determined using experimental and computational approaches. Experimentally, sequence specificity can be determined using techniques as the Protein Binding Microarrays (PBMs), yeast one-hybrid assays, high-throughput sequencing combined with the systematic evolution of ligands by exponential enrichment (ht-SELEX) and DNA-affinity chromatography followed by identification by mass spectrometry (Bulyk 2007; Meng, Brodsky, and Wolfe 2005; Jolma et al. 2010; Tacheny et al. 2013). TFs have a high affinity for their target sequences compared to random genomic DNA. The sequence preference of TFs is identified using computational approaches. Decoding of TF binding sequences starts by obtaining a set of regions that are assumed to be bound by TFs or regulatory regions. Using recent advancements in DNA sequencing, ChIP-seq can be used to identify regions bound by TFs and coactivators. ChIP-seq with TF-specific antibodies can produce a high-resolution map of TF occupancy. Thousands of TFBS affiliated with active enhancers can be predicted using a single TF ChIP-seq experiment. DNase-seq and ATAC-seq are used to identify open chromatin regions, which may correspond to regulatory regions as enhancers or promoters.

Before high-throughput sequencing, methods depended on relatively small sets of sequences (T. L. Bailey and Elkan 1994; Keich and Pevzner 2002; Buhler and Tompa 2002). However, the development of high-throughput techniques assisted the rapid growth in identified regions and accompanied the development of numerous motif discovery methods. Methods relying on de-novo and known motif discovery, aim to reverse-engineer and extract information from regulatory regions based on sequence composition. Known motif discovery methods try to match known motifs in a set of sequences. Results depend on the stringency of the method and methods of scoring. Less rigorous methods allow more mismatches, and therefore result in more matches. De-novo motif discovery is the method looking for overrepresented patterns of nucleotides between a set of sequences of interest and a background set of sequences without any prior knowledge of possible targets. Since nucleotides are not uniformly and randomly distributed, background sequences are necessary to calculate the enrichment of DNA motifs.

Sequence-specific TFs affect gene expression by acting as activators or repressors. Gene regulation is orchestrated by hundreds of TFs interacting with their target genes through cis-regulatory elements in complex GRNs. Having an accurate GRN is an essential step towards understanding gene regulation. The next section focuses on GRNs, different inference methods, and validation approaches.

4. NETWORKS

4.1 Gene regulatory networks

The definition of a network is a collection of distinct elements, nodes, interconnected with each other. Connectivity between nodes is represented by edges, which can be directed or undirected. The first description of a network dates back to the 18th century by the Swiss mathematician, Leonhard Euler. Euler described the Seven Bridges of Königsberg mathematical problem in terms of a graph (network) using nodes and edges to represent the land and bridges (Euler and L 1736). Since then, our understanding of networks and of methods to represent and analyze them has grown dramatically.

Nowadays, networks are found everywhere in our everyday life and can be used to describe systems in every field, including physics, chemistry, social, computers, financial and biology. Representation of systems as networks can provide new approaches for analysis and can provide new insights. Likewise, interactions of TFs with genes, through cis-regulatory elements can be described in Gene Regulatory Networks (GRNs). GRNs allows the systematic representation of regulatory mechanisms, like development or gene regulation, and delineating them is a critical step towards their understanding. They can be undirected or directed, with the edges in undirected networks representing relationship and in the directed networks causation. In the past, building a GRN was based on experimental approaches and primarily focussed on a single TF or gene. Therefore, building a genome-wide map was an expensive and time-consuming process.

The idea of representing gene regulation in networks was initially pioneered by Roy J. Britten and Eric H. Davidson in 1969 and 1971 in sea urchin embryos (Britten and Davidson 1969, 1971). Britten and Davidson presented how genes interact and control products made by other genes in the first model of a gene regulatory network. Their model included cis-regulatory elements and DNA-binding transcription factors before they were experimentally identified at the time, which highlights their pioneering work in the field. They described the interactions of genes as a wiring diagram that illustrates how a gene can influence the transcription of its downstream genes. The diagrams became the standard way to describe networks in their papers.

Nowadays, with computational approaches and advances in high-throughput sequencing, inferring interactions between TF and a gene and therefore a GRN, can be achieved with relatively low cost and time in comparison to the experimental approaches. Combining the vast array of information can lead to a better prediction of GRNs and help in the systematic analysis of regulatory programs (Marbach et al. 2016). Elucidating GRNs can have important

implications for research of health and disease. They can aid in the identification of putative functions for uncharacterized genes and predict expression of target genes (Marbach, Roy, et al. 2012). GRNs can narrow down the potential interactions between TFs and genes, which can then be investigated in a wet lab. Unraveling novel interactions and deciphering of the regulatory program can aid the development of new treatments against congenital and acquired diseases (Hill et al. 2017). They can be used for diagnostic, prognostic and therapeutic purposes to identify subnetworks acting as biomarkers, uncover novel mechanisms involved in cancer and predict key TFs for cellular reprogramming and transdifferentiation between human cell types (Ben-Hamo and Efroni 2011; Dehmer, Mueller, and Emmert-Streib 2013; Rackham et al. 2016). GRNs can provide insights into the molecular mechanism of tissue regeneration and have a significant impact in the area of regenerative medicine (Emmert-Streib, Dehmer, and Haibe-Kains 2014).

4.2 Network properties

In GRNs, the nodes represent genes or TFs and edges convey information regarding a relationship between the two nodes. The networks can be either undirected or directed. In undirected networks, the edge simply indicates associations or functional relationship between the two nodes. In directed networks, the edge indicates a directed relationship between the two nodes. They are used to show a causal effect between two nodes, e.g., a TF regulating a gene and not vice versa. In both types of networks, edges can have a weight which represents the relevance of the connection. Usually, edges with higher weight represent a more reliable connection between the nodes.

Networks have specific properties, which may be used for understanding and exploring the network. One of the properties is the size, which corresponds to the number of nodes that are part of the network. How well these nodes are connected with each other is called the density. Dense networks have highly interconnected nodes. Density can be calculated as the fraction of all possible edges. It also defines its degree at the level of a single node, which corresponds to the number of edges connected to a node. Nodes can have two degrees — the in-degree and out-degree. The in-degree of a node defines the number of incoming edges, while the out-degree specifies the number of outgoing edges. In undirected networks, the in-degree and out-degree are identical. The degree distribution of a network can be constant, random or scale-free. In networks with constant degree distribution, all nodes have an equal amount of connectivity. In random degree distribution, the connectivity of a node is equal to the average connectivity. Finally, scale-free networks do not exhibit any characteristic scale concerning connectivity. Scale-free networks, are a natural result of a preferential attachment process, meaning that as the network grows, new nodes are likely to be connected with other high degree nodes. Another essential property of networks is the structure. Nodes more densely

connected with each other than with the rest can be grouped into communities. Communities can be overlapping or non-overlapping. Identifying possible communities is essential for deciphering the network because they often correspond to different functions and can give a clear and better understanding of how they operate.

GRNs can be modeled using boolean, probabilistic, ordinary differential equations, linear and dynamic models. Many state-of-the-art computational approaches have been proposed for building GRNs. Such methods make use of high-throughput sequencing data, such as RNA expression, chromatin accessibility, histone modification profiles, sequence features, and long-range interactions. Using different approaches and assumptions, these approaches attempt to reverse-engineer GRNs on a genome-wide scale.

4.3 Gene regulatory networks models

4.3.1 Boolean models

Boolean networks are simple and directed graph models, representing nodes as boolean states (active or inactive) and logical relationships (and, or, not) between the nodes (Kaderali and Radde 2008). They are used to represent relationships between TFs and genes, with the presence (active) or absence (inactive) of one of them (Davidich and Bornholdt 2008). They are simple and straightforward but require a large volume of data samples (D'haeseleer, Liang, and Somogyi 2000). Noisy measurements and uncertainties can cause inconsistency (Chai et al. 2014). However, extensions of Boolean Networks, such as Probabilistic Boolean Networks, allow two or more possible transitions to be combined and can cope with uncertainty (Shmulevich et al. 2002).

4.3.2 Probabilistic models

Bayesian and Markov networks belong to the category of probabilistic graphical models. In Bayesian networks, nodes correspond to genes or TFs and they are represented as random variables (Friedman et al. 2000). Edges represent directed probabilistic dependence relations between the nodes, described by the conditional probability distributions. Bayesian networks can be an attractive option for modeling GRNs because they can deal with noisy measurements and missing data (de Jong 2002; Kaderali and Radde 2008). With a set of incomplete measurements, the model can successfully predict the topology of a network. Since they were initially proposed in 2000, they have become an increasingly popular method for inferring GRNs (Friedman et al. 2000; Liu et al. 2016; Hartemink et al. 2001; Pe'er et al. 2001). Nonetheless, they have high computational complexity, therefore are not preferred for large-scale networks, and because they are acyclic graphs, they cannot model feedback loops. Extensions of Bayesian Networks, such as the Dynamic Bayesian Networks, can model feedback loops only in time-series data (de Jong 2002; Baba et al. 2014).

4.3.3 Ordinary differential equations models

Ordinary differential equations (ODEs) models are the most widespread method for modeling or simulating dynamic systems using a set of ODEs (de Jong 2002). They describe relationships between an independent variable and its derivatives. Unlike probabilistic models, ODEs are a deterministic approach and interactions between genes represent causation and not statistical dependency. In GRNs, they model the dynamic change of gene expression as a function of a variable, such as time-series mRNA genes levels of related genes. However, a complete characterization of the model needs a lot of prior information to specify the values of different model parameters, which makes ODEs models only feasible for small networks (Ko, Voit, and Wang 2009; Morris et al. 2010; F. M. Khan et al. 2014). For more extensive networks, it becomes challenging and computationally intensive to estimate all the parameters.

4.4 Gene regulatory networks inference

4.4.1 Expression-based inference approaches

Expression-based approaches use expression patterns to infer regulatory interactions. These approaches are based on correlation, information theory or feature selection. In correlation-based or co-expression networks, nodes represent genes and the edge the similarity in expression profiles between them. The rationale behind this is that genes with similar expression can be functionally related. The similarity between a pair of genes is calculated using a pairwise similarity function, such as the Pearson correlation coefficient. If the similarity score is greater than the selected threshold, then the two genes are connected. Co-expression networks have low computational complexity, but because the edges are based on correlation, they are undirected and only assume that there is a functional relationship between the pair of genes. Examples of correlation-based methods are the MutualRank and Z-score. The MutualRank method calculates and ranks the correlation between each pair of genes (Obayashi and Kinoshita 2009). The z-score method uses wild-type and knockout data to calculate the expression differences between genes and its targets (Prill et al. 2010). The idea behind it is that in the knockout experiments the affected genes are the ones regulated by the corresponding TF. The most common approach to analyze co-expression networks is with clustering (Eisen et al. 1998; Serin et al. 2016).

Information theory approaches aim to capture more statistical dependencies from the expression data using a method called mutual information. Mutual information is a generation of the pairwise correlation coefficient and measures the degree of mutual dependence between two genes. Examples of mutual information approaches are RelNet, CLR, ARACNE, C3NET, PCIT, and Relevance Networks. The RelNet is based on a method called relevance networks which was initially proposed in 1999 by Butte and Kohane (Butte and Kohane 1999). For each gene pair the mutual information is calculated and if it is above the threshold an

edge is drawn (Butte and Kohane 2000). The CLR (Context Likelihood or Relatedness network) method estimates the mutual information between two genes as the correlation coefficient from their expression. To eliminate noise, it takes into account the background distribution and only pairs who deviate more are more likely to be interacting (Faith et al. 2007). The ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks), is the most famous among the mutual information methods (Margolin et al. 2006; Basso et al. 2005). It is based on a similar hypothesis but can differentiate between direct and indirect edges. Firstly, using the Gaussian kernel density method it computes the dependency between two genes. Then using the Data Processing Inequality, it tries to reduce the number of false positives by removing the weakest edge in every triplet set of genes. The PCIT (Partial Correlation coefficient with Information Theory) is using partial correlation coefficient together with mutual information to calculate the dependencies between two genes. Similar to ARACNE, PCIT uses gene triplets to remove indirect interactions.

Feature selection approaches try to select a relevant subset of features to build a model and select the true regulators for each gene (Bellot et al. 2015). It reduces the search space by integrating prior knowledge and removing no-TFs genes. Examples of feature selection methods are GENIE3 and MRNET. The GENIE3 (GEne Network Inference with Ensemble of trees) approach is based on the Random Forests algorithm and feature selection. Using a gene-by-gene approach it predicts the expression of a target gene from the input data (Huynh-Thu et al. 2010).

4.4.2 Inference of gene regulatory networks using regulatory data

Conventional GRN inference approaches aim to infer interactions based on expression data. As discussed previously, those approaches find it difficult to distinguish direct from indirect interactions. It is known that gene regulation is influenced by a diverse range of elements, such as transcription factor binding, chromatin accessibility and three-dimensional organization of the genome. Hence, taking into account this information can aid in better understanding of gene regulation. Numerous computational approaches have been emerging that combine these measurements to infer GRNs and link TFs to their target genes. Combining all this information leads to better and more accurate networks (Marbach, Roy, et al. 2012). Identifying TF binding can be achieved using DNA sequencing technologies. Genome-wide maps histone modification and chromatin accessibility profiles can be used for the genome-wide mapping of regulatory regions of genes, such as promoters and enhancers, and using computational methods identify TFs binding. However, this is limited to TFs with known binding sites. Using experimental approaches, such as ChIP-seq to identify TF-bound regions, followed by motif discovery can overcome this limitation. Such approaches have been used in the development of GRNs in human and *Drosophila* and it was shown that they are more accurate in predicting

known edges than expression-based networks (Neph et al. 2012; Marbach et al. 2016; Marbach, Roy, et al. 2012).

4.4.3 Machine learning data integration

Machine learning is a sub-discipline of computer science which aims to design algorithms that help computerized systems to "learn" from observed data and subsequently identify patterns, make decisions or predictions. Machine learning-based data integration approaches to infer GRNs can be divided into three categories; supervised, unsupervised and semi-supervised (Libbrecht and Noble 2015).

Supervised approaches rely on exploiting known information. Using prior biological knowledge, they aim to predict new regulatory interactions. Algorithms are trained on a predefined set of regulatory interactions and infer features information based on multiple resources, such as expression, binding sites, and ChIP-seq signal. This information is then used to determine and classify new TF-gene interactions. If a gene has common information between the known targets in the training set is classified as a true target, otherwise as false. Such methods depend to a high degree on the quality of the training set and, therefore, are limited to only well-studied organisms with many known, interactions. Supervised algorithms can be separated into two main categories - classification and regression. Classification algorithms aim to classify the data into specific classes, for example, positive or negative interactions between as TF and a gene. Supervised methods include Bayesian classifiers, Support Vector Machines, Random Forests, Neural Networks, and regression models. Support Vector Machines are a binary classification method (Cortes and Vapnik 1995). Examples of supervised GRN inference approaches include GENIE3, SIRENE, and Beacon (Huynh-Thu et al. 2010; Mordelet and Vert 2008; Ni et al. 2016).

Semi-supervised learning approaches are hybrid approaches falling halfway between supervised and unsupervised (Chapelle, Schölkopf, and Zien 2006). They are a blend of both approaches. In addition to a significant amount of unlabeled data, they take advantage of any prior information (labeled data). Often in GRN inference approaches, the unlabeled data can be the gene expression or TF binding and prior information TF to target gene interactions. GRN inference semi-supervised approaches include SEREND (Ernst et al. 2008).

Unsupervised approaches try to infer conclusions from unlabeled data. They tend to rely on expression and binding data and unlike supervised methods, data are not classified into positive and negative interactions. They do not require to be trained on training examples; therefore they are less prone to overfitting and can be used in less studied organisms where gene regulation information is sparse. Unsupervised methods include CLR, ARACNE, WGCNA,

TIGRESS, and MRNET (Faith et al. 2007; Margolin et al. 2006; Langfelder and Horvath 2008; Haury et al. 2012; Meyer et al. 2007).

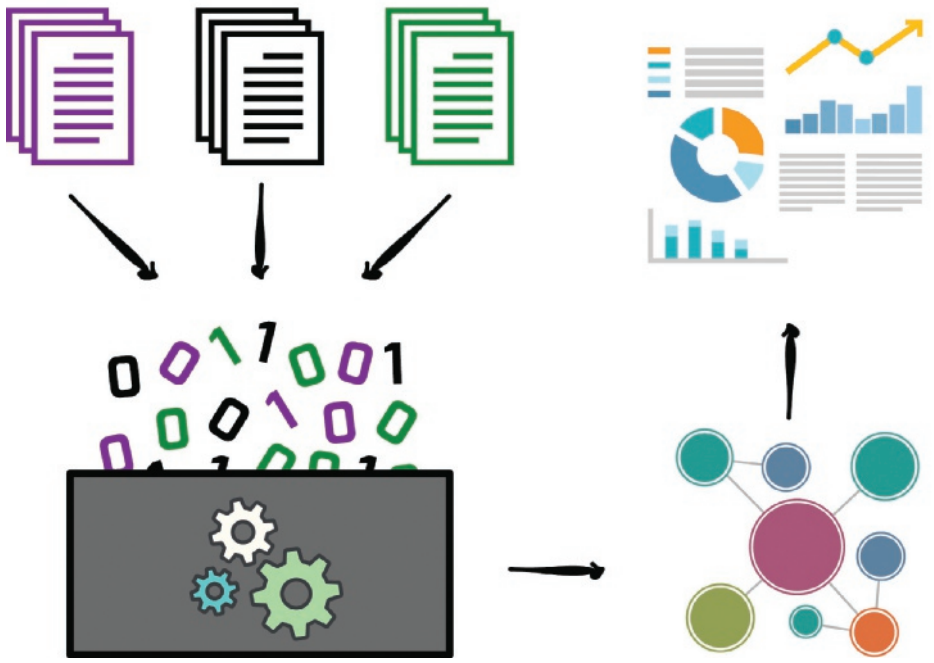


Figure 3. A typical workflow for the inference and analysis of a gene regulatory network. Expression or other regulatory data are used as input for different computational approaches. The resulted networks are then validated and analyzed to infer conclusions.

4.4.4 Ensemble models

Sir Francis Galton, a British statistician, described the first known instance of using the collective information to get more accurate results (Galton 1907). In his publication in 1907, Galton described a weight guessing competition at a festival in Cornwall. He asked the visitors to guess the weight of an ox, but none of the visitors' guesses were correct. However, he observed that when he averaged their guesses, the answer was always almost close to the actual weight. In 2004, James Surowiecki in his book "The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations" argues that using information from different sources can benefit decision making in business and economics fields. This method eliminates limitations in knowledge or potential bias. In machine learning, this approach is called ensemble learning. Likewise, information from multiple "experts" is combined in a single model and used to obtain a better prediction.

In the fifth systems biology challenge of DREAM (Dialogue for Reverse Engineering Assessments and Methods) project they compared the performance of 35 different inference methods among several datasets (Marbach, Costello, et al. 2012). They discovered that while some of the individual methods performed well in some datasets, none performed consistently well among all the datasets. Methods based their predictions on different hypotheses, therefore, they have their advantages and limitations. By combining all predictions into a single model, they observed that methods could complement each other by reducing the limitations and improving the final predictions.

4.5 Validation

Validation is an essential and necessary part of the epistemology of GRN inference. Inferential and scientific ability of GRN inference approaches need to be evaluated to estimate what extent their predictions represent reality.

Inferential validation relates to the predicting power of the approach to infer a network (Dougherty 2011). The predicted network is compared with a known network, which is generally referred to as the gold standard. A gold standard is often constructed from experimentally-validated direct regulatory interactions obtained from the literature which are treated as True Positives (TP). Commonly used for statistical validation is the Area Under Curve (AUC) of Receiver Operator Characteristic (ROC) and Precision-Recall (PR) curves (Golicher et al. 2012; Manning, Manning, and Schütze 1999). ROC and PR curves show the performance of a binary classifier in predicting true or false interactions. ROC evaluates the approach by comparing the number of correctly classified interactions (True Positive Rate (TPR)) against the number of incorrectly classified interactions (False Positive Rate (FPR)). For ROC an AUC of 0.5 suggests the performance is no better than random and an AUC of 1.0 represents a perfect prediction. PR compares the precision of the approach with the recall. Precision is the fraction of TP over the total number of positive predictions (TP + False Positive (FP)). The recall is the fraction of TP against the total of TP and False Negative (FN) interactions.

As mentioned earlier, supervised and semi-supervised approaches use prior biological knowledge, also known as the training set, to predict new regulatory interactions. Concerning validation, part of the training set is used as the validation set. This technique is called cross-validation. Validation sets usually consist of a randomly selected 30-40% of the data, with the remaining used as the training set. However, this technique can lead to overfitting. To handle overfitting, this process is repeated k times, for k-fold validation, and the performance of the method is measured as the mean of the corresponding AUCs.

The quality of an inferred network can also be evaluated on its properties and the biological relevance of predicted interactions. Studies have shown that in- and out-degrees of GRNs follow a power-law distribution, therefore, predicted networks degrees are expected to follow the same distribution (Guelzim et al. 2002; Balázsi and Oltvai 2005; Borotkanics and Lehmann 2015). Based on the idea that genes regulated by similar TFs tend to have a similar function or being expressed in similar tissues, the biological relevance of predicted interactions can be verified with ontology annotation (Marbach, Roy, et al. 2012). Genes sharing same regulators are expected to have a similar function and therefore having significantly higher enrichment in the same Gene Ontology (GO) functional annotation terms compared to randomly-generated networks. In the same vein, co-regulated genes are likely to be enriched in specific Anatomical Ontology (AO) terms. This approach can be used to verify the whole network per se, along with its structure. Densely connected nodes forming communities often include targets sharing the same regulators, having the same function or being expressed in the same tissues.

Scientific validation relates to the ability to make observations from the predicted network (Dougherty 2011). Network predictions, such as interactions, are compared with experimental observations. Using a TF gene knockout experiment to inactivate a TF will disturb its binding to the cis-regulatory regions and affect the expression of its real target genes. Comparing the expression of the predicted targets in the knockout experiment to the wild type can be a clear indication if the prediction was correct. Alternatives to knockout include knockdown and overexpression for decreasing or increasing the expression of the TF.

5. OVERVIEW OF THE THESIS

Embryonic development is a highly dynamic process orchestrated by large and complex GRNs of hundreds of TFs interacting with chromatin and genes through distant enhancers. Deciphering and understanding these interactions can have an immense impact on human health. It can contribute to the understanding of vertebrate genes and the study of functional regulation and consequently assist in drug discovery and therefore help against congenital diseases.

What is the state of chromatin at a specific developmental time and in a specific part of the genome? Which TFs are responsible for the activation of genes? The rise of high-throughput sequencing technology allowed for experiments that helped to answer these and many more questions. NGS made possible the genome-wide profiling of histone modifications and TF binding at relatively low cost. Having a detailed map of histone modifications can help us identify which are the chromatin states and how they change during development. TF

binding data aid the development of GRNs and therefore strengthen our understanding of gene regulation.

This thesis focuses on histone modifications, enhancer and GRNs dynamics during early embryonic development and in evolution.

Chapter Two focuses on histone modification dynamics during embryonic development in *X. tropicalis* embryos. Using ChIP-seq data, we generated epigenome reference maps and identified chromatin states based on overlapping histone modifications. States were divided into seven groups; Polycomb, poised enhancers, active enhancers, transcribed regions, promoters, heterochromatin, and unmodified regions. We showed that active and repressive marks are dynamic during development. Finally, we find that the deposition of H3K4me3 and H3K27me3 histone modifications is mainly determined by maternal factors, while recruitment of p300 to enhancers is regulated by zygotic factors. In this chapter, I was involved in generating the epigenome maps, performed the analysis of chromatin state dynamics and supported the other analyses.

Chapter Three describes the development of fluff, a software package that allows for simple exploration, clustering, and visualization of high-throughput sequencing data mapped to a reference genome. In this chapter, we illustrate the functionality of fluff to identify spatial and dynamic patterns of histone modifications. Using DNase I hypersensitive sites in H1 human embryonic stem cells differentiated into mesenchymal, mesendoderm, neuronal progenitor and trophoblast lineages, we identified clusters specific to those lineages. For this chapter, I wrote the code, analyzed the data, prepared the figures and wrote the manuscript.

In **Chapter Four** we study the regulatory innovations that contributed to the genomic evolution of this *X. laevis* and the immediate effects of hybridization. We studied subgenome-specific enhancers and found them to be enriched for transposable elements carrying TF binding sites. To study the early regulatory remodeling events following hybridization, we generated *X. tropicalis* × *X. laevis* hybrid embryos. We found that young and active *X. tropicalis* DNA transposons are responsible for the recruitment of p300 in hybrid embryos. In this chapter, I was involved in drafting the manuscript, designed the analysis, performed genome alignment and analyses of differentially methylated regions and hybrids.

Chapter Five focuses on the dynamics of gene-regulatory networks during embryonic development. We describe a novel ensemble method for inferring GRNs. The method integrates binding of the p300 (Ep300) coactivator, transcription factor expression and transcription factor motifs to infer gene-regulatory interactions. We applied the method to genome-wide datasets

available in *X. tropicalis* embryos during different developmental stages. We identified stage-specific network communities associated with Gene Ontology and *Xenopus* Anatomy Ontology terms. For each TF we assigned an influence score based on its differentially expressed targets and we identified key TFs for each developmental stage. Finally, we constructed spatial networks for the animal cap, vegetal mass, ventral, lateral and dorsal marginal zones. In this chapter, I designed and performed the analyses, wrote the code, prepared the figures and wrote the manuscript.

Finally, **Chapter Six** summarizes the work and findings described in this thesis and discusses future novel work to advance our understanding of gene regulation.

REFERENCES

Allfrey, V. G., R. Faulkner, and A. E. Mirsky. 1964. "ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS." Proceedings of the National Academy of Sciences of the United States of America 51 (May): 786–94.

Avery, O. T., C. M. Macleod, and M. McCarty. 1944. "STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III." The Journal of Experimental Medicine 79 (2): 137–58.

Baba, N., M. S. Mohamad, A. H. Mohamed Salleh, M. H. Ahmad Hijazi, L. E. Chai, M. M. Zainuddin, and S. Deris. 2014. "Continuous Dynamic Bayesian Network for Gene Regulatory Network Modelling." In 2014 International Conference on Computational Science and Technology (ICCST), 1–5.

Bailey, Timothy L., and Charles Elkan. 1995. "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization." Machine Learning 21 (1): 51–80.

Bailey, T. L., and C. Elkan. 1994. "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers." Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology 2: 28–36.

Balázsi, Gábor, and Zoltán N. Oltvai. 2005. "Sensing Your Surroundings: How Transcription-Regulatory Networks of the Cell Discern Environmental Signals." Science's STKE: Signal Transduction Knowledge Environment 2005 (282): e20.

Banerji, J., S. Rusconi, and W. Schaffner. 1981. "Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences." Cell 27 (2 Pt 1): 299–308.

Bannister, Andrew J., and Tony Kouzarides. 2011. "Regulation of Chromatin by Histone Modifications." Cell Research 21 (3): 381–95.

Basso, Katia, Adam A. Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. 2005. "Reverse Engineering of Regulatory Networks in Human B Cells." Nature Genetics 37 (4): 382–90.

Bastle, Ryan M., and Ian S. Maze. 2019. "Chromatin Regulation in Complex Brain Disorders." Current Opinion in Behavioral Sciences 25 (February): 57–65.

Bellot, Pau, Catharina Olsen, Philippe Salembier, Albert Oliveras-Vergés, and Patrick E. Meyer. 2015. "NetBenchmark: A Bioconductor Package for Reproducible Benchmarks of Gene Regulatory Network Inference." BMC Bioinformatics 16 (September): 312.

Ben-Hamo, Rotem, and Sol Efroni. 2011. "Gene Expression and Network-Based Analysis Reveals a Novel Role for Hsa-miR-9 and Drug Control over the p38 Network in Glioblastoma Multiforme Progression." Genome Medicine 3 (11): 77.

Benos, Panayiotis V., Martha L. Bulyk, and Gary D. Stormo. 2002. "Additivity in Protein-DNA Interactions: How Good an Approximation Is It?" Nucleic Acids Research 30 (20): 4442–51.

Berman, Benjamin P., Yutaka Nibu, Barret D. Pfeiffer, Pavel Tomancak, Susan E. Celniker, Michael Levine, Gerald M. Rubin, and Michael B. Eisen. 2002. "Exploiting Transcription Factor Binding Site Clustering to Identify Cis-Regulatory Modules Involved in Pattern Formation in the Drosophila Genome." Proceedings of the National Academy of Sciences of the United States of America 99 (2): 757–62.

Blackwood, E. M., and J. T. Kadonaga. 1998. "Going the Distance: A Current View of Enhancer Action." Science 281 (5373): 60–63.

Blitz, Ira L., Gregor Andelfinger, and Marko E. Horb. 2006. "Germ Layers to Organs: Using *Xenopus* to Study 'Later' Development." Seminars in Cell & Developmental Biology 17 (1): 133–45.

Bonn, Stefan, Robert P. Zinzen, Charles Girardot, E. Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen E. M. Furlong. 2012. "Tissue-Specific Analysis of Chromatin State Identifies Temporal Signatures of Enhancer Activity during Embryonic Development." Nature Genetics 44 (2): 148–56.

Borneman, Anthony R., Tara A. Gianoulis, Zhengdong D. Zhang, Haiyuan Yu, Joel Rozowsky, Michael R. Seringhaus, Lu Yong Wang, Mark Gerstein, and Michael Snyder. 2007. "Divergence of Transcription Factor Binding Sites across Related Yeast Species." Science 317 (5839): 815–19.

Borotkanics, Robert, and Harold Lehmann. 2015. "Network Motifs That Recur across Species, Including Gene Regulatory and Protein-Protein Interaction Networks." Archives of Toxicology 89 (4): 489–99.

Bradley, Robert K., Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A. Tonkin, Mark D. Biggin, and Michael B. Eisen. 2010. "Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species." PLoS Biology 8 (3): e1000343.

Britten, R. J., and E. H. Davidson. 1969. "Gene Regulation for Higher Cells: A Theory." Science 165 (3891): 349–57.

———. 1971. "Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty." The Quarterly Review of Biology 46 (2): 111–38.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.] 109 (January): 21.29.1–9.

Buhler, Jeremy, and Martin Tompa. 2002. "Finding Motifs Using Random Projections." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 9 (2): 225–42.

Bulyk, Martha L. 2007. "Protein Binding Microarrays for the Characterization of DNA-Protein Interactions." Advances in Biochemical Engineering/biotechnology 104: 65–85.

Bulyk, Martha L., Philip L. F. Johnson, and George M. Church. 2002. "Nucleotides of Transcription Factor Binding Sites Exert Interdependent Effects on the Binding Affinities of Transcription Factors." Nucleic Acids Research 30 (5): 1255–61.

Butte, A. J., and I. S. Kohane. 1999. "MUTUAL INFORMATION RELEVANCE NETWORKS: FUNCTIONAL GENOMIC CLUSTERING USING PAIRWISE ENTROPY MEASUREMENTS." In Biocomputing 2000, 418–29. WORLD SCIENTIFIC.

———. 2000. "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 418–29.

Calo, Eliezer, and Joanna Wysocka. 2013. "Modification of Enhancer Chromatin: What, How, and Why?" Molecular Cell 49 (5): 825–37.

Cannavò, Enrico, Pierre Khoueiry, David A. Garfield, Paul Geeleher, Thomas Zichner, E. Hilary Gustafson, Lucia Ciglar, Jan O. Korbel, and Eileen E. M. Furlong. 2016. "Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks." Current Biology: CB 26 (1): 38–51.

Chai, Lian En, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. 2014. "A Review on the Computational Approaches for Gene Regulatory Network Construction." Computers in Biology and Medicine 48 (May): 55–65.

Chan, H. M., and N. B. La Thangue. 2001. "p300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds." Journal of Cell Science 114 (Pt 13): 2363–73.

Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien. 2006. Semi-Supervised Learning. MIT Press.

Chatterjee, Sumantra, and Nadav Ahituv. 2017. "Gene Regulatory Elements, Major Drivers of Human Disease." Annual Review of Genomics and Human Genetics 18 (August): 45–63.

Chrun, Emanuely Silva, Filipe Modolo, and Filipe Ivan Daniel. 2017. "Histone Modifications: A Review about the Presence of This Epigenetic Phenomenon in Carcinogenesis." Pathology, Research and Practice 213 (11): 1329–39.

Cockerill, Peter N. 2011. "Structure and Function of Active Chromatin and DNase I Hypersensitive Sites." The FEBS Journal 278 (13): 2182–2210.

Cornish-Bowden, A. 1985. "Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984." Nucleic Acids Research 13 (9): 3021–30.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." Machine Learning 20 (3): 273–97.

Crease, D. J., S. Dyson, and J. B. Gurdon. 1998. "Cooperation between the Activin and Wnt Pathways in the Spatial Control of Organizer Gene Expression." Proceedings of the National Academy of Sciences of the United States of America 95 (8): 4398–4403.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State." Proceedings of the National Academy of Sciences of the United States of America 107 (50): 21931–36.

Davidich, Maria I., and Stefan Bornholdt. 2008. "Boolean Network Model Predicts Cell Cycle Sequence of Fission Yeast." PloS One 3 (2): e1672.

Dehmer, Matthias, Laurin A. J. Mueller, and Frank Emmert-Streib. 2013. "Quantitative Network Measures as Biomarkers for Classifying Prostate Cancer Disease States: A Systems Approach to Diagnostic Biomarkers." PloS One 8 (11): e77602.

Del Bene, Filippo, Laurence Ettwiller, Dorota Skowronska-Krawczyk, Herwig Baier, Jean-Marc Matter, Ewan Birney, and Joachim Wittbrodt. 2007. "In Vivo Validation of a Computationally Predicted Conserved Ath5 Target Gene Set." PLoS Genetics 3 (9): 1661–71.

D'haeseleer, P., S. Liang, and R. Somogyi. 2000. "Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering." Bioinformatics 16 (8): 707–26.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." Nature 485 (7398): 376–80.

Dougherty, Edward R. 2011. "Validation of Gene Regulatory Networks: Scientific and Inferential." Briefings in Bioinformatics 12 (3): 245–52.

Drouin, Jacques. 2016. "Epigenetic Mechanisms of Pituitary Cell Fate Specification." In Stem Cells in Neuroendocrinology, edited by Donald Pfaff and Yves Christen. Cham (CH): Springer.

Eckner, R., Z. Arany, M. Ewen, W. Sellers, and D. M. Livingston. 1994. "The Adenovirus E1A-Associated 300-kD Protein Exhibits Properties of a Transcriptional Coactivator and Belongs to an Evolutionarily Conserved Family." Cold Spring Harbor Symposia on Quantitative Biology 59: 85–95.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. "Cluster Analysis and Display of Genome-Wide Expression Patterns." Proceedings of the National Academy of Sciences of the United States of America 95 (25): 14863–68.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains. 2014. "Gene Regulatory Networks and Their Applications: Understanding Biological and Medical Problems in Terms of Networks." Frontiers in Cell and Developmental Biology 2 (August): 1–7.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." Nature 489 (7414): 57–74.

Ernst, Jason, Qasim K. Beg, Krin A. Kay, Gábor Balázsi, Zoltán N. Oltvai, and Ziv Bar-Joseph. 2008. "A Semi-Supervised Method for Predicting Transcription Factor-Gene Interactions in Escherichia Coli." PLoS Computational Biology 4 (3): e1000044.

Ernst, Jason, and Manolis Kellis. 2012. "ChromHMM: Automating Chromatin-State Discovery and Characterization." Nature Methods 9 (3): 215–16.

———. 2017. "Chromatin-State Discovery and Genome Annotation with ChromHMM." Nature Protocols 12 (12): 2478–92.

Euler, and L. 1736. "Solutio Problematis Ad Geometriam Situs Pertinensis." Comm. Acad. Sci. Imper. Petropol. 8: 128–40.

Faith, Jeremiah J., Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. 2007. "Large-Scale Mapping and Validation of Escherichia Coli Transcriptional Regulation from a Compendium of Expression Profiles." PLoS Biology 5 (1): e8.

Festuccia, Nicola, Rodrigo Osorno, Florian Halbritter, Violetta Karwacki-Neisius, Pablo Navarro, Douglas Colby, Frederick Wong, Adam Yates, Simon R. Tomlinson, and Ian Chambers. 2012. "Esrrb Is a Direct Nanog Target Gene That Can Substitute for Nanog Function in Pluripotent Cells." Cell Stem Cell 11 (4): 477–90.

Flemming, Walther. 1882. Zellsubstanz, Kern Und Zelltheilung. Vogel.

Foat, Barrett C., Alexandre V. Morozov, and Harmen J. Bussemaker. 2006. "Statistical Mechanical Modeling of Genome-Wide Transcription Factor Occupancy Data by MatrixREDUCE." Bioinformatics 22 (14): e141–49.

Fortin, Connor H., Katharina V. Schulze, and Gregory A. Babbitt. 2015. "TRX-LOGOS - a Graphical Tool to Demonstrate DNA Information Content Dependent upon Backbone Dynamics in Addition to Base Sequence." Source Code for Biology and Medicine 10 (September): 10.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er. 2000. "Using Bayesian Networks to Analyze Expression Data." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 7 (3-4): 601–20.

Galton, Francis. 1907. "Vox Populi." Nature 75 (March): 450.

Gilbert, Scott F. 2000. Developmental Biology. Palgrave Macmillan.

Gillette, Thomas G., and Joseph A. Hill. 2015. "Readers, Writers, and Erasers: Chromatin as the Whiteboard of Heart Disease." Circulation Research 116 (7): 1245–53.

Golicher, Duncan, Andrew Ford, Luis Cayuela, and Adrian Newton. 2012. "Pseudo-Absences, Pseudo-Models and Pseudo-Niches: Pitfalls of Model Selection Based on the Area under the Curve." International Journal of Geographical Information Science: IJGIS 26 (11): 2049–63.

Gross, D. S., and W. T. Garrard. 1988. "Nuclease Hypersensitive Sites in Chromatin." Annual Review of Biochemistry 57: 159–97.

Guelzim, Nabil, Samuele Bottani, Paul Bourgine, and François Képès. 2002. "Topological and Causal Structure of the Yeast Transcriptional Regulatory Network." Nature Genetics 31 (1): 60–63.

Hartemink, A. J., D. K. Gifford, T. S. Jaakkola, and R. A. Young. 2001. "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 422–33.

Haury, Anne-Claire, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. 2012. "TIGRESS: Trustful Inference of Gene REgulation Using Stability Selection." BMC Systems Biology 6 (November): 145.

Hay, Deborah, Jim R. Hughes, Christian Babbs, James O. J. Davies, Bryony J. Graham, Lars Hanssen, Mira T. Kassouf, et al. 2016. "Genetic Dissection of the α-Globin Super-Enhancer in Vivo." Nature Genetics 48 (8): 895–903.

He, Bin Z., Alisha K. Holloway, Sebastian J. Maerkl, and Martin Kreitman. 2011. "Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules." PLoS Genetics 7 (4): e1002053.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. "Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression." Nature 459 (7243): 108–12.

Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome." Nature Genetics 39 (3): 311–18.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." Molecular Cell 38 (4): 576–89.

Hempel, Annemarie, and Michael Kühl. 2016. "A Matter of the Heart: The African Clawed Frog *Xenopus* as a Model for Studying Vertebrate Cardiogenesis and Congenital Heart Defects." Journal of Cardiovascular Development and Disease 3 (2). https://doi.org/10.3390/jcdd3020021.

Hershey, A. D., and Martha Chase. 1952. "INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE." The Journal of General Physiology 36 (1): 39–56.

Hill, Jonathon T., Bradley Demarest, Bushra Gorsi, Megan Smith, and H. Joseph Yost. 2017. "Heart Morphogenesis Gene Regulatory Networks Revealed by Temporal Expression Analysis." Development 144 (19): 3487–98.

Hoffman, Michael M., Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. 2012. "Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation." Nature Methods 9 (5): 473–76.

Hong, Joung-Woo, David A. Hendrix, and Michael S. Levine. 2008. "Shadow Enhancers as a Source of Evolutionary Novelty." Science 321 (5894): 1314.

Huang, Jialiang, Kailong Li, Wenqing Cai, Xin Liu, Yuannyu Zhang, Stuart H. Orkin, Jian Xu, and Guo-Cheng Yuan. 2018. "Dissecting Super-Enhancer Hierarchy Based on Chromatin Interactions." Nature Communications 9 (1): 943.

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. 2010. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods." PloS One 5 (9). https://doi.org/10.1371/journal.pone.0012776.

Iwafuchi-Doi, Makiko, and Kenneth S. Zaret. 2014. "Pioneer Transcription Factors in Cell Reprogramming." Genes & Development 28 (24): 2679–92.

Jin, Qihuang, Li-Rong Yu, Lifeng Wang, Zhijing Zhang, Lawryn H. Kasper, Ji-Eun Lee, Chaochen Wang, Paul K. Brindle, Sharon Y. R. Dent, and Kai Ge. 2011. "Distinct Roles of GCN5/PCAF-Mediated H3K9ac and CBP/p300-Mediated H3K18/27ac in Nuclear Receptor Transactivation." The EMBO Journal 30 (2): 249–62.

Jolma, Arttu, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, et al. 2010. "Multiplexed Massively Parallel SELEX for Characterization of Human Transcription Factor Binding Specificities." Genome Research 20 (6): 861–73.

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. "DNA-Binding Specificities of Human Transcription Factors." Cell 152 (1-2): 327–39.

Jong, Hidde de. 2002. "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 9 (1): 67–103.

Kaderali, Lars, and Nicole Radde. 2008. "Inferring Gene Regulatory Networks from Expression Data." In Computational Intelligence in Bioinformatics, edited by Arpad Kelemen, Ajith Abraham, and Yuehui Chen, 33–74. Berlin, Heidelberg: Springer Berlin Heidelberg.

Keich, U., and P. A. Pevzner. 2002. "Finding Motifs in the Twilight Zone." Bioinformatics 18 (10): 1374–81.

Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, et al. 2018. "JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework." Nucleic Acids Research 46 (D1): D260–66.

Khan, Faiz M., Ulf Schmitz, Svetoslav Nikolov, David Engelmann, Brigitte M. Pützer, Olaf Wolkenhauer, and Julio Vera. 2014. "Hybrid Modeling of the Crosstalk between Signaling and Transcriptional Networks Using Ordinary Differential Equations and Multi-Valued Logic." Biochimica et Biophysica Acta 1844 (1 Pt B): 289–98.

Kheradpour, Pouya, Alexander Stark, Sushmita Roy, and Manolis Kellis. 2007. "Reliable Prediction of Regulator Targets Using 12 Drosophila Genomes." Genome Research 17 (12): 1919–31.

Kiecker, Clemens, Thomas Bates, and Esther Bell. 2016. "Molecular Specification of Germ Layers in Vertebrate Embryos." Cellular and Molecular Life Sciences: CMLS 73 (5): 923–47.

Kim, T. K., T. Lagrange, Y. H. Wang, J. D. Griffith, D. Reinberg, and R. H. Ebright. 1997. "Trajectory of DNA in the RNA Polymerase II Transcription Preinitiation Complex." Proceedings of the National Academy of Sciences of the United States of America 94 (23): 12268–73.

Ko, Chih-Lung, Eberhard O. Voit, and Feng-Sheng Wang. 2009. "Estimating Parameters for Generalized Mass Action Models with Connectivity Information." BMC Bioinformatics 10 (May): 140.

Kornberg, R. D., and J. O. Thomas. 1974. "Chromatin Structure; Oligomers of the Histones." Science 184 (4139): 865–68.

Kornberg, Roger D. 2007. "The Molecular Basis of Eukaryotic Transcription." Proceedings of the National Academy of Sciences of the United States of America 104 (32): 12955–61.

Kouzarides, Tony. 2007. "Chromatin Modifications and Their Function." Cell 128 (4): 693–705.

Kribelbauer, Judith F., Oleg Laptenko, Siying Chen, Gabriella D. Martini, William A. Freed-Pastor, Carol Prives, Richard S. Mann, and Harmen J. Bussemaker. 2017. "Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes." Cell Reports 19 (11): 2383–95.

Laat, Wouter de, and Denis Duboule. 2013. "Topology of Mammalian Developmental Enhancers and Their Regulatory Landscapes." Nature 502 (7472): 499–506.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. "The Human Transcription Factors." Cell 172 (4): 650–65.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." BMC Bioinformatics 9 (December): 559.

Lee, T. I., and R. A. Young. 2000. "Transcription of Eukaryotic Protein-Coding Genes." Annual Review of Genetics 34: 77–137.

Leung, H. C. M., and F. Y. L. Chin. n.d. "An Efficient Algorithm for the Extended (l, D)-Motif Problem with Unknown Number of Binding Sites." In Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), 11–18. IEEE.

Leung, Henry C. M., Francis Y. L. Chin, and Bethany M. Y. Chan. 2007. "Discovering Motifs with Transcription Factor Domain Knowledge." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 472–83.

Levine, Michael, and Robert Tjian. 2003. "Transcription Regulation and Animal Diversity." Nature 424 (6945): 147–51.

Libbrecht, Maxwell W., and William Stafford Noble. 2015. "Machine Learning Applications in Genetics and Genomics." Nature Reviews. Genetics 16 (6): 321–32.

Li, Guoliang, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, et al. 2012. "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation." Cell 148 (1-2): 84–98.

Liu, Fei, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen. 2016. "Inference of Gene Regulatory Network Based on Local Bayesian Networks." PLoS Computational Biology 12 (8): e1005024.

Lovén, Jakob, Heather A. Hoke, Charles Y. Lin, Ashley Lau, David A. Orlando, Christopher R. Vakoc, James E. Bradner, Tong Ihn Lee, and Richard A. Young. 2013. "Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers." Cell 153 (2): 320–34.

Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. 1997. "Crystal Structure of the Nucleosome Core Particle at 2.8 A Resolution." Nature 389 (6648): 251–60.

Lupien, Mathieu, Jérôme Eeckhoute, Clifford A. Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S. Carroll, X. Shirley Liu, and Myles Brown. 2008. "FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription." Cell 132 (6): 958–70.

Manning, Christopher D., Christopher D. Manning, and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Marbach, Daniel, James C. Costello, Robert Küffner, Nicci Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, et al. 2012. "Wisdom of Crowds for Robust Gene Network Inference." Nature Methods 9 (8): 796–804.

Marbach, Daniel, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. 2016. "Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations across Complex Diseases." Nature Methods 13 (4): 366–70.

Marbach, Daniel, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher a. Bristow, and Manolis Kellis. 2012. "Predictive Regulatory Models in Drosophila Melanogaster by Integrative Inference of Transcriptional Networks." Genome Research 22 (7): 1334–49.

Margolin, Adam A., Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context." BMC Bioinformatics 7 Suppl 1 (March): S7.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. "Transcriptional Regulatory Elements in the Human Genome." Annual Review of Genomics and Human Genetics 7 (1): 29–59.

Matsui, T., J. Segall, P. A. Weil, and R. G. Roeder. 1980. "Multiple Factors Required for Accurate Initiation of Transcription by Purified RNA Polymerase II." The Journal of Biological Chemistry 255 (24): 11992–96.

Meng, Xiangdong, Michael H. Brodsky, and Scot A. Wolfe. 2005. "A Bacterial One-Hybrid System for Determining the DNA-Binding Specificity of Transcription Factors." Nature Biotechnology 23 (8): 988–94.

Meyer, Patrick E., Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. 2007. "Information-Theoretic Inference of Large Transcriptional Regulatory Networks." EURASIP Journal on Bioinformatics & Systems Biology, 79879.

Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, V. Afzal, H. Hadeiba, K. Shinkai, E. M. Rubin, and R. M. Locksley. 2001. "Deletion of a Coordinate Regulator of Type 2 Cytokine Expression in Mice." Nature Immunology 2 (9): 842–47.

Moorthy, Sakthi D., Scott Davidson, Virlana M. Shchuka, Gurdeep Singh, Nakisa Malek-Gilani, Lida Langroudi, Alexandre Martchenko, Vincent So, Neil N. Macpherson, and Jennifer A. Mitchell. 2017. "Enhancers and Super-Enhancers Have an Equivalent Regulatory Role in Embryonic Stem Cells through Regulation of Single or Multiple Genes." Genome Research 27 (2): 246–58.

Mordelet, Fantine, and Jean-Philippe Vert. 2008. "SIRENE: Supervised Inference of Regulatory Networks." Bioinformatics 24 (16): i76–82.

Morris, Melody K., Julio Saez-Rodriguez, Peter K. Sorger, and Douglas A. Lauffenburger. 2010. "Logic-Based Models for the Analysis of Cell Signaling Networks." Biochemistry 49 (15): 3216–24.

Murakami, Kenji, Kuang-Lei Tsai, Nir Kalisman, David A. Bushnell, Francisco J. Asturias, and Roger D. Kornberg. 2015. "Structure of an RNA Polymerase II Preinitiation Complex." Proceedings of the National Academy of Sciences of the United States of America 112 (44): 13543–48.

Nakato, Ryuichiro, and Katsuhiko Shirahige. 2017. "Recent Advances in ChIP-Seq Analysis: From Quality Management to Whole-Genome Annotation." Briefings in Bioinformatics 18 (2): 279–90.

Neph, Shane, Andrew B. Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A. Stamatoyannopoulos. 2012. "Circuitry and Dynamics of Human Transcription Factor Regulatory Networks." Cell 150 (6): 1274–86.

Niederriter, Adrienne R., Arushi Varshney, Stephen C. J. Parker, and Donna M. Martin. 2015. "Super Enhancers in Cancers, Complex Disease, and Developmental Disorders." Genes 6 (4): 1183–1200.

Nishibuchi, Gohei, and Jérôme Déjardin. 2017. "The Molecular Basis of the Organization of Repetitive DNA-Containing Constitutive Heterochromatin in Mammals." Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology 25 (1): 77–87.

Ni, Ying, Delasa Aghamirzaie, Haitham Elmarakeby, Eva Collakova, Song Li, Ruth Grene, and Lenwood S. Heath. 2016. "A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis." Frontiers in Plant Science 7 (December): 1936.

Obayashi, Takeshi, and Kengo Kinoshita. 2009. "Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression." DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes 16 (5): 249–60.

Ostuni, Renato, Viviana Piccolo, Iros Barozzi, Sara Polletti, Alberto Termanini, Silvia Bonifacio, Alessia Curina, Elena Prosperini, Serena Ghisletti, and Gioacchino Natoli. 2013. "Latent Enhancers Activated by Stimulation in Differentiated Cells." Cell 152 (1-2): 157–71.

Parker, Stephen C. J., Michael L. Stitzel, D. Leland Taylor, Jose Miguel Orozco, Michael R. Erdos, Jennifer A. Akiyama, Kelly Lammerts van Bueren, et al. 2013. "Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants." Proceedings of the National Academy of Sciences of the United States of America 110 (44): 17921–26.

Pasini, Diego, Martina Malatesta, Hye Ryung Jung, Julian Walfridsson, Anton Willer, Linda Olsson, Julie Skotte, et al. 2010. "Characterization of an Antagonistic Switch between Histone H3 Lysine 27 Methylation and Acetylation in the Transcriptional Regulation of Polycomb Group Target Genes." Nucleic Acids Research 38 (15): 4958–69.

Pe'er, D., A. Regev, G. Elidan, and N. Friedman. 2001. "Inferring Subnetworks from Perturbed Expression Profiles." Bioinformatics 17 Suppl 1: S215–24.

Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. "Enhancers: Five Essential Questions." Nature Reviews. Genetics 14 (4): 288–95.

Perry, Michael W., Alistair N. Boettiger, Jacques P. Bothma, and Michael Levine. 2010. "Shadow Enhancers Foster Robustness of Drosophila Gastrulation." Current Biology: CB 20 (17): 1562–67.

Pesole, G., N. Prunella, S. Liuni, M. Attimonelli, and C. Saccone. 1992. "WORDUP: An Efficient Algorithm for Discovering Statistically Significant Patterns in DNA Sequences." Nucleic Acids Research 20 (11): 2871–75.

Pombo, Ana, and Niall Dillon. 2015. "Three-Dimensional Genome Architecture: Players and Mechanisms." Nature Reviews. Molecular Cell Biology 16 (4): 245–57.

Pott, Sebastian, and Jason D. Lieb. 2015. "What Are Super-Enhancers?" Nature Genetics 47 (1): 8–12.

Prill, Robert J., Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. 2010. "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges." PloS One 5 (2): e9202.

Rackham, Owen J. L., Jaber Firas, Hai Fang, Matt E. Oates, Melissa L. Holmes, Anja S. Knaupp, FANTOM Consortium, et al. 2016. "A Predictive Computational Framework for Direct Reprogramming between Human Cell Types." Nature Genetics 48 (3): 331–35.

Rada-Iglesias, Alvaro, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. 2011. "A Unique Chromatin Signature Uncovers Early Developmental Enhancers in Humans." Nature 470 (7333): 279–83.

Richmond, T. J., J. T. Finch, B. Rushton, D. Rhodes, and A. Klug. 1984. "Structure of the Nucleosome Core Particle at 7 A Resolution." Nature 311 (5986): 532–37.

Rohs, Remo, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. 2009. "The Role of DNA Shape in Protein-DNA Recognition." Nature 461 (7268): 1248–53.

Schmidt, Dominic, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, et al. 2010. "Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding." Science 328 (5981): 1036–40.

Schmitt, Stefan M., Mazhar Gull, and André W. Brändli. 2014. "Engineering *Xenopus* Embryos for Phenotypic Drug Discovery Screening." Advanced Drug Delivery Reviews 69-70 (April): 225–46.

Schneider, T. D., and R. M. Stephens. 1990. "Sequence Logos: A New Way to Display Consensus Sequences." Nucleic Acids Research 18 (20): 6097–6100.

Serin, Elise A. R., Harm Nijveen, Henk W. M. Hilhorst, and Wilco Ligterink. 2016. "Learning from Co-Expression Networks: Possibilities and Challenges." Frontiers in Plant Science 7 (April): 1–18.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. "Transcriptional Enhancers: From Properties to Genome-Wide Predictions." Nature Reviews. Genetics 15 (4): 272–86.

Shmulevich, Ilya, Edward R. Dougherty, Seungchan Kim, and Wei Zhang. 2002. "Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks." Bioinformatics 18 (2): 261–74.

Sinha, Saurabh. 2003. "Discriminative Motifs." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 10 (3-4): 599–615.

Sinha, S., and M. Tompa. 2000. "A Statistical Method for Finding Transcription Factor Binding Sites." Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology 8: 344–54.

Song, Lingyun, and Gregory E. Crawford. 2010. "DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells." Cold Spring Harbor Protocols 2010 (2): db.prot5384.

Spemann, H., and Hilde Mangold. 1924. "über Induktion von Embryonalanlagen Durch Implantation Artfremder Organisatoren." Archiv Für Mikroskopische Anatomie Und Entwicklungsmechanik 100 (3): 599–638.

Spitz, François, and Eileen E. M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." Nature Reviews. Genetics 13 (9): 613–26.

Stender, J. D., K. Kim, T. H. Charn, B. Komm, K. C. N. Chang, W. L. Kraus, C. Benner, C. K. Glass, and B. S. Katzenellenbogen. 2010. "Genome-Wide Analysis of Estrogen Receptor DNA Binding and Tethering Mechanisms Identifies Runx1 as a Novel Tethering Factor in Receptor-Mediated Transcriptional Activation." Molecular and Cellular Biology 30 (16): 3943–55.

Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht. 1982. "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in *E. Coli*." Nucleic Acids Research 10 (9): 2997–3011.

Struhl, K. 1995. "Yeast Transcriptional Regulatory Mechanisms." Annual Review of Genetics 29: 651–74.

Tacheny, A., M. Dieu, T. Arnould, and P. Renard. 2013. "Mass Spectrometry-Based Identification of Proteins Interacting with Nucleic Acids." Journal of Proteomics 94 (December): 89–109.

Tuğrul, Murat, Tiago Paixão, Nicholas H. Barton, and Gašper Tkačik. 2015. "Dynamics of Transcription Factor Binding Site Evolution." PLoS Genetics 11 (11): e1005639.

Wang, Qianben, Jason S. Carroll, and Myles Brown. 2005. "Spatial and Temporal Recruitment of Androgen Receptor and Its Coactivators Involves Chromosomal Looping and Polymerase Tracking." Molecular Cell 19 (5): 631–42.

Wang, Shu-Ping, Zhanyun Tang, Chun-Wei Chen, Miho Shimada, Richard P. Koche, Lan-Hsin Wang, Tomoyoshi Nakadai, et al. 2017. "A UTX-MLL4-p300 Transcriptional Regulatory Network Coordinately Shapes Active Enhancer Landscapes for Eliciting Transcription." Molecular Cell 67 (2): 308–21.e6.

Weirauch, Matthew T., Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, et al. 2013. "Evaluation of Methods for Modeling Transcription Factor Sequence Specificity." Nature Biotechnology 31 (2): 126–34.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity." Cell 158 (6): 1431–43.

Wheeler, Grant N., and André W. Brändli. 2009. "Simple Vertebrate Models for Chemical Genetics and Drug Discovery Screens: Lessons from Zebrafish and *Xenopus*." Developmental Dynamics: An Official Publication of the American Association of Anatomists 238 (6): 1287–1308.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." Cell 153 (2): 307–19.

Wolpert, Lewis, Cheryll Tickle, and Alfonso Martinez Arias. 2015. Principles of Development. Oxford University Press.

Xie, Shiqi, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. 2017. "Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells." Molecular Cell 66 (2): 285–99.e5.

Xu, Yang, Shuang Zhang, Shaofeng Lin, Yaping Guo, Wankun Deng, Ying Zhang, and Yu Xue. 2017. "WERAM: A Database of Writers, Erasers and Readers of Histone Acetylation and Methylation in Eukaryotes." Nucleic Acids Research 45 (D1): D264–70.

Yang, Lin, Yaron Orenstein, Arttu Jolma, Yimeng Yin, Jussi Taipale, Ron Shamir, and Remo Rohs. 2017. "Transcription Factor Family-Specific DNA Shape Readout Revealed by Quantitative Specificity Models." Molecular Systems Biology 13 (2): 910.

Yin, Yimeng, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, et al. 2017. "Impact of Cytosine Methylation on DNA Binding Specificities of Human Transcription Factors." Science 356 (6337). https://doi.org/10.1126/science.aaj2239.

Zabidi, Muhammad A., Cosmas D. Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. 2015. "Enhancer-Core-Promoter Specificity Separates Developmental and Housekeeping Gene Regulation." Nature 518 (7540): 556–59.

Zaret, Kenneth S., and Jason S. Carroll. 2011. "Pioneer Transcription Factors: Establishing Competence for Gene Expression." Genes & Development 25 (21): 2227–41.

Zawel, L., and D. Reinberg. 1993. "Initiation of Transcription by RNA Polymerase II: A Multi-Step Process." Progress in Nucleic Acid Research and Molecular Biology 44: 67–108.

CHAPTER TWO

Embryonic transcription is controlled by maternally defined chromatin state

Saartje Hontelez
Ila van Kruijsbergen
Georgios Georgiou
Simon J. van Heeringen
Ozren Bogdanovic
Ryan Lister
Gert Jan C. Veenstra

ABSTRACT

Histone-modifying enzymes are required for cell identity and lineage commitment, however little is known about the regulatory origins of the epigenome during embryonic development. Here we generate a comprehensive set of epigenome reference maps, which we use to determine the extent to which maternal factors shape chromatin state in *Xenopus* embryos. Using α-amanitin to inhibit zygotic transcription, we find that the majority of H3K4me3- and H3K27me3-enriched regions form a maternally defined epigenetic regulatory space with an underlying logic of hypomethylated islands. This maternal regulatory space extends to a substantial proportion of neurula stage-activated promoters. In contrast, p300 recruitment to distal regulatory regions requires embryonic transcription at most loci. The results show that H3K4me3 and H3K27me3 are part of a regulatory space that exerts an extended maternal control well into post-gastrulation development and highlight the combinatorial action of maternal and zygotic factors through proximal and distal regulatory sequences.

INTRODUCTION

During early embryonic development cells differentiate, acquiring specific transcription and protein expression profiles. Histone modifications can control the activity of genes through regulatory elements in a cell-type-specific manner [1-4]. Recent advances have been made in the annotation of functional genomic elements of mammalian cells, *Drosophila* and *Caenorhabditis* through genome-wide profiling of chromatin marks[5, 6]. Immediately after fertilization, the embryonic genome is transcriptionally silent, and zygotic genome activation (ZGA) occurs after a number of mitotic cycles [7]. In *Drosophila* and zebrafish (*Danio rerio*) ZGA starts after 8 and 9 mitotic cycles, respectively, in mammals transcription starts at the two-cell stage[8, 9], whereas in *Xenopus* this happens after the first 12 cleavages at the mid-blastula transition (MBT)[10-12]. Permissive H3K4me3 and repressive H3K27me3 histone modifications emerge during blastula and gastrula stages[13-16]. To date, little is known about the origin and specification of the epigenome in embryonic development of vertebrates, which is essential for understanding physiological cell lineage commitment and differentiation.

To explore the developmental origins of epigenetic regulation we have generated epigenome reference maps during early development of *Xenopus tropicalis* embryos and assessed the need for embryonic transcription in their acquisition. We find a hierarchical appearance of histone modifications, with a priority for promoter marks which are deposited hours before transcription activation on regions with hypomethylated DNA. Surprisingly, the promoter H3K4me3 and the Polycomb H3K27me3 modifications are largely maternally defined (MaD), providing maternal epigenetic control of gene activation that extends well into neurula and tailbud stages. By contrast, p300 recruitment to distal regulatory elements is largely under the control of zygotic factors. Moreover, this maternal-proximal and zygotic-distal dichotomy of gene regulatory sequences also differentiates between early and late Wnt signalling target genes, suggesting that different levels of permissiveness are involved in temporal target gene selection.

RESULTS

Progressive specification of chromatin state

We have performed chromatin immunoprecipitation (ChIP) sequencing of eight histone modifications, RNA polymerase II (RNAPII) and the enhancer protein p300 at five stages of development: blastula (st. 9), gastrula (st. 10.5, 12.5), neurula (st. 16) and tailbud (st. 30). These experiments allow identification of enhancers (H3K4me1, p300)[17-20], promoters (H3K4me3, H3K9ac)[14, 21-23], transcribed regions (H3K36me3, RNAPII)[22] and repressed and heterochromatic domains (H3K27me3, H3K9me2, H3K9me3 and H4K20me3)[1, 14, 24, 25]. In addition we generated

pre-MBT (st. 8) maps for three histone modifications (H3K4me3, H3K9ac and H3K27me3) and single-base resolution DNA methylome maps using whole-genome bisulfite sequencing of blastula and gastrula (st. 9 and 10.5) embryos (Fig. 1; Supplementary Fig. 1).

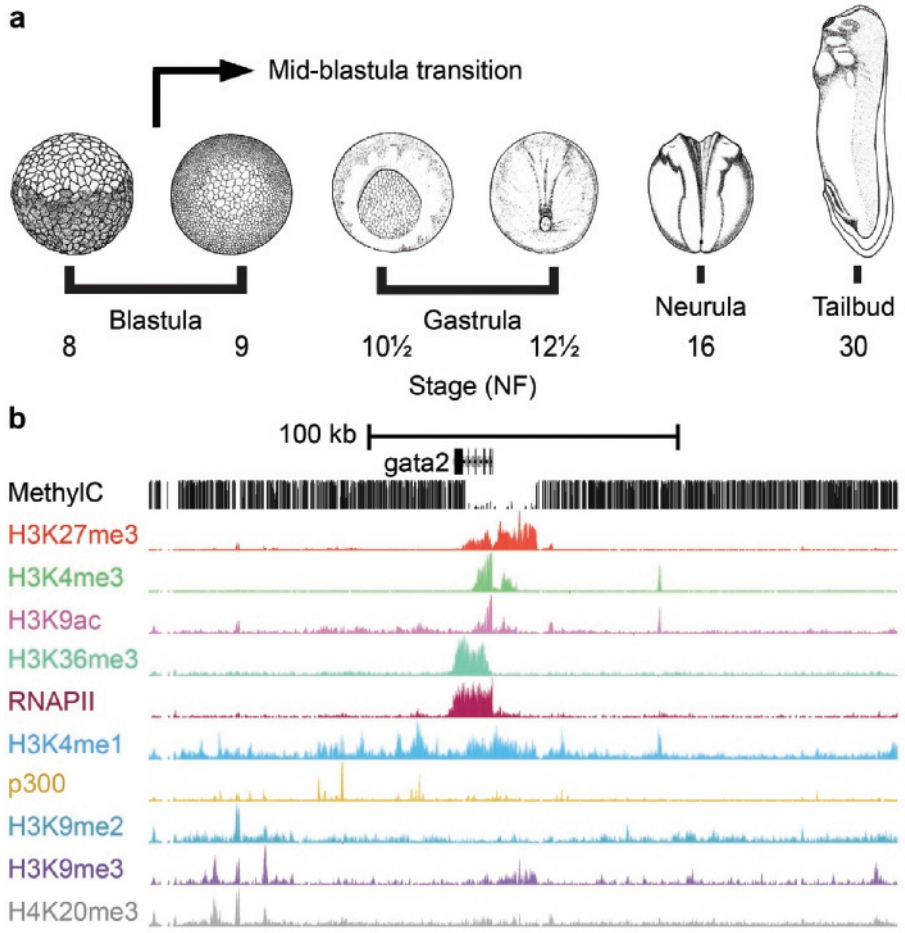


Figure 1. Reference epigenome maps of Xenopus tropicalis development. (a) Genome-wide profiles were generated for stages 8 and 9 (blastula, before and after MBT), 10.5 and 12.5 (gastrula), 16 (neurula) and 30 (tailbud). Adapted from Tan, M.H. et al. Genome Res. 23, 201–216 (2013), under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/ by/3.0/. **(b)** *Gata2* locus with late gastrula (stage 10.5) methylC-seq, ChIP- seq enrichment of histone modifications, RNAPII and p300 (cf. Supplementary Figs 1 and 2).

Our data set consists of 2.7 billion aligned sequence reads representing the most comprehensive set of epigenome reference maps of vertebrate embryos to date. Using a Hidden Markov Model approach[26] we have identified 19 chromatin states based on co-occurring

ChIP signals (Fig. 2a). This analysis identifies combinations of ChIP signals at specific genomic sequences without distinguishing between overlapping histone modifications that result from regional or cell-type specificity and co-occurrence in the same cells[14]. Seven main groups were recognized, namely (i) Polycomb (H3K27me3, deposited by Polycomb Repressive Complex 2 (PRC2)), (ii) poised enhancers, (iii) p300-bound enhancers, (iv) transcribed regions, (v) promoters, (vi) heterochromatin and (vii) unmodified regions (Fig. 2a; Supplementary Fig. 2). Alluvial plots of state coverage per stage show that all states increase in coverage during development, except for the unmodified state (Fig. 2b, Supplementary Fig. 2a). Unmodified regions decrease in coverage during development, however, even at tailbud stage 67% of the total epigenome remains naive for the modifications and bound proteins in our data set (Supplementary Fig. 2b). Promoter coverage remains relatively constant during development from blastula to tailbud stages, in contrast to the Polycomb state which increases in coverage during gastrulation. P300-bound enhancers are highly dynamic during development (Fig. 2b). Global enrichment levels of modified regions show similar dynamics, and reveal a priority for promoter marking at or before the blastula stage, followed by enhancer activation and heterochromatic repression during late blastula and gastrulation stages (Supplementary Fig. 3a,b). A detailed time course between fertilization and early gastrulation shows that both H3K4me3 and H3K9ac emerge hours before the start of embryonic transcription (Supplementary Fig. 3c). We and others have previously reported that H3K4me3 is acquired during blastula stages[14]. Indeed, H3K4me3 and H3K9ac levels increase strongly before the MBT, well before embryonic transcription starts. This however raises the question to what extent histone modifications are regulated by maternal or embryonic factors.

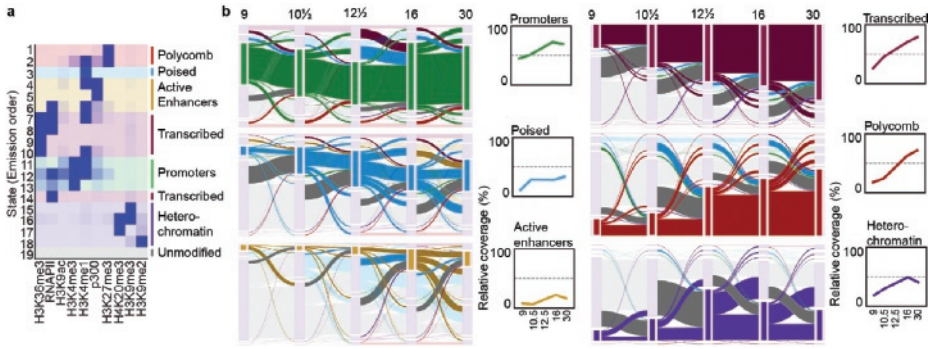


Figure 2. Chromatin state dynamics. (a) Emission states (same for all developmental stages) of the hidden Markov model, identifying the 19 most prevalent combinations of histone modifications and bound proteins. From top to bottom: Polycomb (red), Poised enhancers and promoters (blue), Active Enhancers (gold), Transcribed (dark magenta), Promoter (green), Heterochromatin (purple) and unmodified (grey). **(b)** Alluvial plots of chromatin state coverage during development. Each plot shows the transitions (to and from the highlighted group of chromatin states) across developmental stages (stages 9–30). The height represents the base pair coverage of the chromatin state relative to the modified genome. The 'modified genome' has a chromatin state other than unmodified in any of the stages 9–30. From top to bottom left: promoters (green), poised (blue), p300-bound enhancers (gold). From top to bottom right: transcribed (dark magenta), Polycomb (red) and heterochromatin (purple). Line plots: Chromatin state coverage per stage as a percentage of the modified genome.

Maternal and zygotic epigenetic regulation

To determine the maternal and zygotic contributions to chromatin state, we used α-amanitin to block embryonic transcription (Fig. 3a). α-Amanitin blocks the translocation of RNA polymerase II (RNAPII) on DNA, thereby preventing transcript elongation[27]. It is therefore expected that injection of α-amanitin into embryos will stall RNAPII, immobilizing it on DNA after its recruitment to pre-initiation complexes. Indeed, both RNAPII elongation and embryonic transcription were effectively blocked in α-amanitin-injected embryos (Fig. 3b, c; Supplementary Fig. 4a). New transcription is necessary for gastrulation[11, 28, 29], but α-amanitin-injected embryos survive to the equivalent of stage 11 control embryos. ChIP-sequencing of replicates of α-amanitin- injected and control embryos (stage 11) revealed that the majority of H3K4me3 (86%) and H3K27me3 (90%) regions are consistently modified with these modifications independently of embryonic transcription (Fig. 3d; Supplementary Fig. 4b, c). This is especially surprising given the temporal hierarchy of H3K27me3 and H3K4me3, and the relatively late acquisition of H3K27me3 (Fig. 2b). By contrast, only 15% of the p300-bound regions recruit p300 independently of active transcription (Fig. 3d). This suggests that the promoter-permissive H3K4me3 mark and the Polycomb-repressive H3K27me3 mark are mostly controlled by maternal factors (maternally defined, MaD), whereas p300 binding to regulatory regions is largely zygotically defined (ZyD). Regions with MaD H3K4me3 and H3K27me3 acquire

these modifications more robustly and also earlier during development compared with ZyD regions (Supplementary Fig. 4d). By contrast, ZyD p300-bound regions show more robust p300 recruitment during gastrulation compared with p300 MaD regions. These data show a pervasive maternal influence on the developmental acquisition of key histone modifications.

DNA methylation logic of maternal control

Trimethylation of H3K4 and H3K27 has been associated with CpG density and a lack of DNA methylation. The Set1 and related MLL complexes are responsible for H3K4me3[10]. Set1 is recruited to hypomethylated CpG domains via the Cxxc1 protein (Cfp1)[30-32]. In the absence of H3K4me3, PRC2 binding to hypomethylated CpGs results in H3K27me3 and inhibition of gene activation[13, 33]. Using our whole-genome bisulfite sequencing data we determined that MaD H3K4me3 promoters are predominantly hypomethylated (Fig. 4a; Supplementary Fig. 5a; Supplementary Data 1). Conversely, promoters decorated with ZyD H3K4me3 almost exclusively have highly methylated promoters. Demethylation of ZyD promoters was not detected, and methylation levels of MaD and ZyD regions were similar in stage 9 and stage 10.5 (Supplementary Fig. 5a, b). In addition, H3K4me3 often extends asymmetrically from promoters into gene bodies (+1-2 kb from transcription start site (TSS)); (Supplementary Fig. 5c), likely representing the second and third nucleosomes that are trimethylated via RNAPII-recruited Set1 in actively transcribed genes[34]. Concordantly, α-amanitin reduces H3K4me3 at downstream positions. Interestingly, we also find poised enhancers that gain H3K4me3 in α-amanitin-injected embryos and which exhibit intermediate to high levels of DNA methylation (Supplementary Fig. 5d, e).

The majority of promoters with ZyD H3K27me3 shows intermediate to high levels of DNA methylation (Fig. 4a; Supplementary Fig. 5a; Supplementary Data 1). Some of the MaD H3K27me3 regions are methylated, but the highly enriched H3K27me3 domains (larger dots) are almost exclusively both maternally defined and hypomethylated. This is illustrated by the hoxd cluster which harbours a large hypomethylated domain with MaD H3K4me3 and H3K27me3 (Fig. 4b). There are also examples of reciprocal changes of H3K4 and H3K27 methylation, for example at the hypermethylated promoters of nodal1 and nodal2.

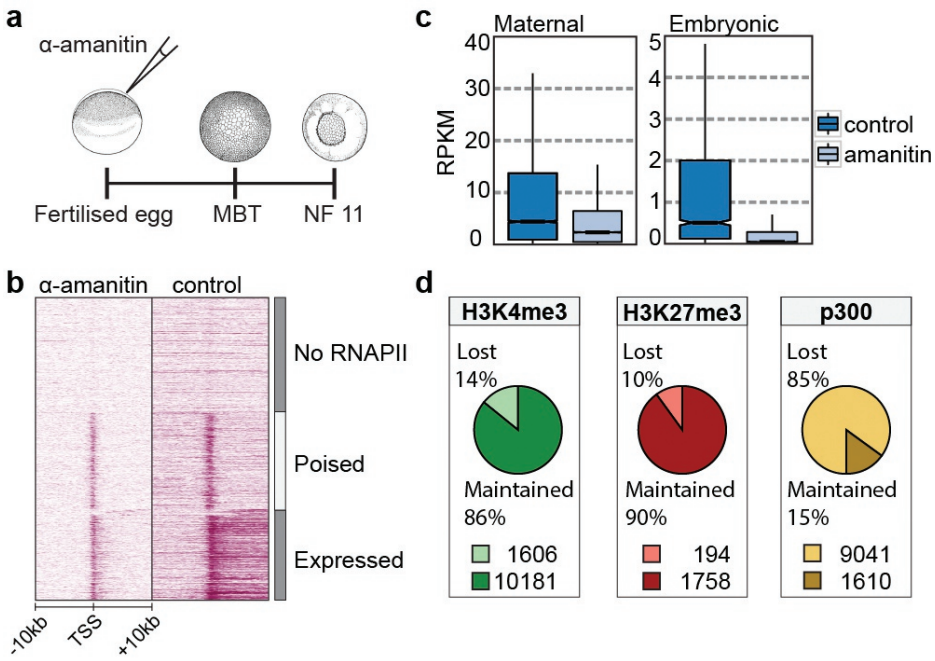


Figure 3. Developmental acquisition of chromatin states. (a) Inhibition of embryonic transcription with a-amanitin, adapted from Tan, M.H. et al. Genome Res. 23, 201–216 (2013), under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http:// creativecommons.org/licenses/by/3.0/. **(b)** RNAPII on the TSS of genes in control and a-amanitin-injected embryos (stage 11). **(c)** Box plots showing RNA expression levels (RPKM) of maternal and embryonic transcribed genes in control and a-amanitin-injected embryos (stage 11). Box: 25th (bottom), 50th (internal band), 75th (top) percentiles. Whiskers: 1.5 × interquartile range of the lower and upper quartiles, respectively. **(d)** ChIP-sequencing on chromatin of a-amanitin-injected and control embryos reveals maternal and zygotic origins of H3K4me3, H3K27me3 or p300 binding. Data from two biological replicates, see Supplementary fig. 4.

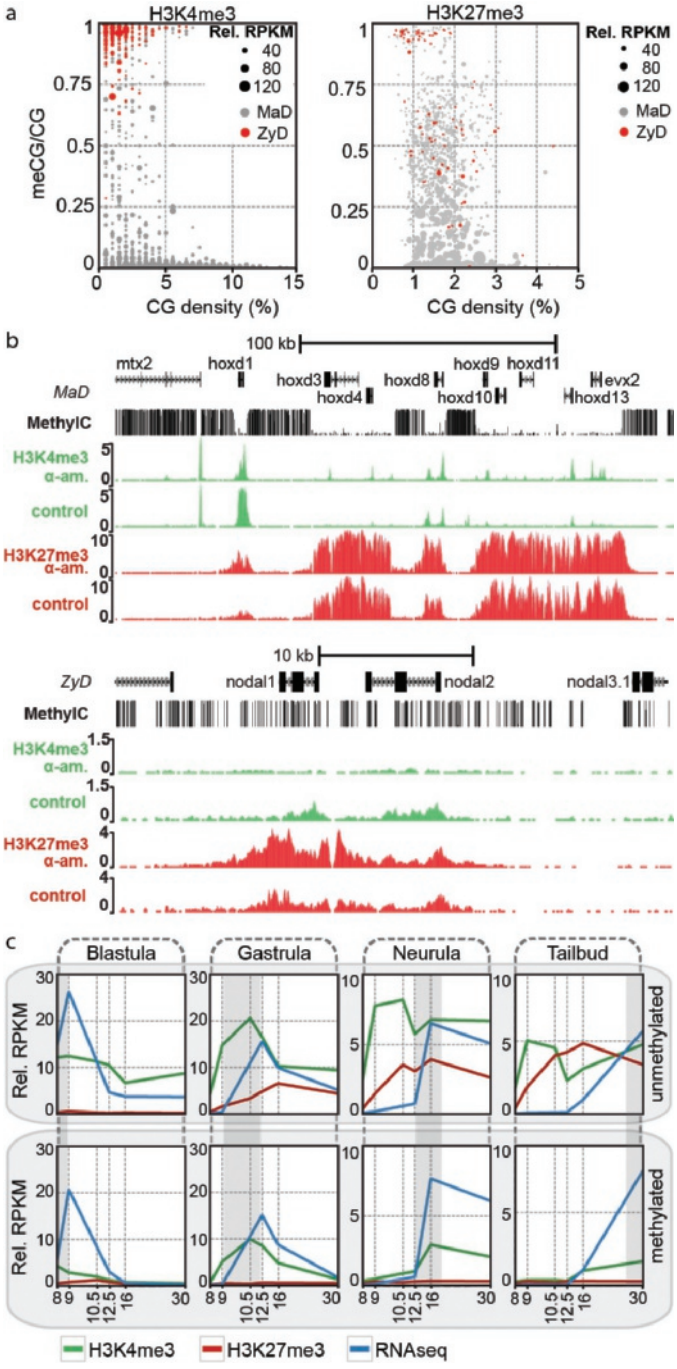


Figure 4. DNA methylation logic of maternally versus zygotically defined H3K4me3 and H3K27me3. (a) CpG density and methylation at stage 9 of promoters (H3K4me3: ±100 bp from TSS; H3K27me3: ±2.5 kb from TSS) that contain a zygotic defined (ZyD, lost in a-amanitin treated embryos,

red) or maternal defined (MaD, maintained in a-amanitin treated embryos, grey) peak for H3K4me3 (left) or H3K27me3 (right) after inhibition of embryonic transcription. The size of the dot indicates the relative RPKM of the histone modification (background corrected). **(b)** Hoxd (MaD) and nodal1, -2 (ZyD) loci with stage 9 methylC-seq, H3K4me3 and H3K27me3 in control and a-amanitin-injected embryos. **(c)** Developmental profiles of H3K4me3 and H3K27me3 (median background corrected RPKM) at genes without detectable maternal mRNA do correlate with activation for methylated promoters (lower panels) but not for hypomethylated CpG island promoters (upper panels).

ZyD p300-bound regions are generally hypermethylated, whereas MaD p300-bound regions show a variable degree of DNA methylation (Supplementary Fig. 5e). However, promoters that overlap with MaD p300 peaks are hypomethylated in 77% of the cases, whereas 96% of the promoters that are associated with ZyD p300 peaks are hypermethylated (Supplementary Fig. 5f), showing that p300-recruiting hypomethylated promoters tend to be under complete maternal control, for both H3K4 methylation and p300 recruitment.

To further explore the relationships between DNA methylation, histone modifications and developmental activation of transcription we determined correlations with different measures of gene activity such as RNA-seq and ChIP-seq of RNAPII and H3K36me3 (Supplementary Fig. 6). We find that H3K36me3 and RNAPII in gene bodies correlate well with each other but less with transcript levels (RNA-seq), presumably due to the effects of RNA stability. A much lower correlation was found between either measure of gene activity and the promoter marks H3K4me3 and H3K9ac, especially at early stages. In part this may be caused by time delays of transcriptional activation relative to acquisition of permissive histone modifications[14, 15]. It raises the question to what extent a lack of DNA methylation at promoters, which is associated with MaD H3K4me3, uncouples promoter marking and transcriptional activation. Therefore, we grouped transcribed genes without detectable maternal messenger RNA[35] based on the stage of maximum expression and DNA methylation (Fig. 4c). We find that developmentally activated promoters with hypomethylated CpG islands are trimethylated at H3K4 or H3K27 early on, irrespective of the time of transcriptional activation. By contrast, methylated promoters show a much closer relation between H3K4me3 and gene expression. Although H3K4me3 is known to stabilize the transcription initiation factor Taf3 (a subunit of TFIID) and can also interact with the chromatin remodeller Chd1[36-38], hypomethylated promoters gain H3K4me3 autonomously with their hypomethylated CpG island status, independent of embryonic transcription.

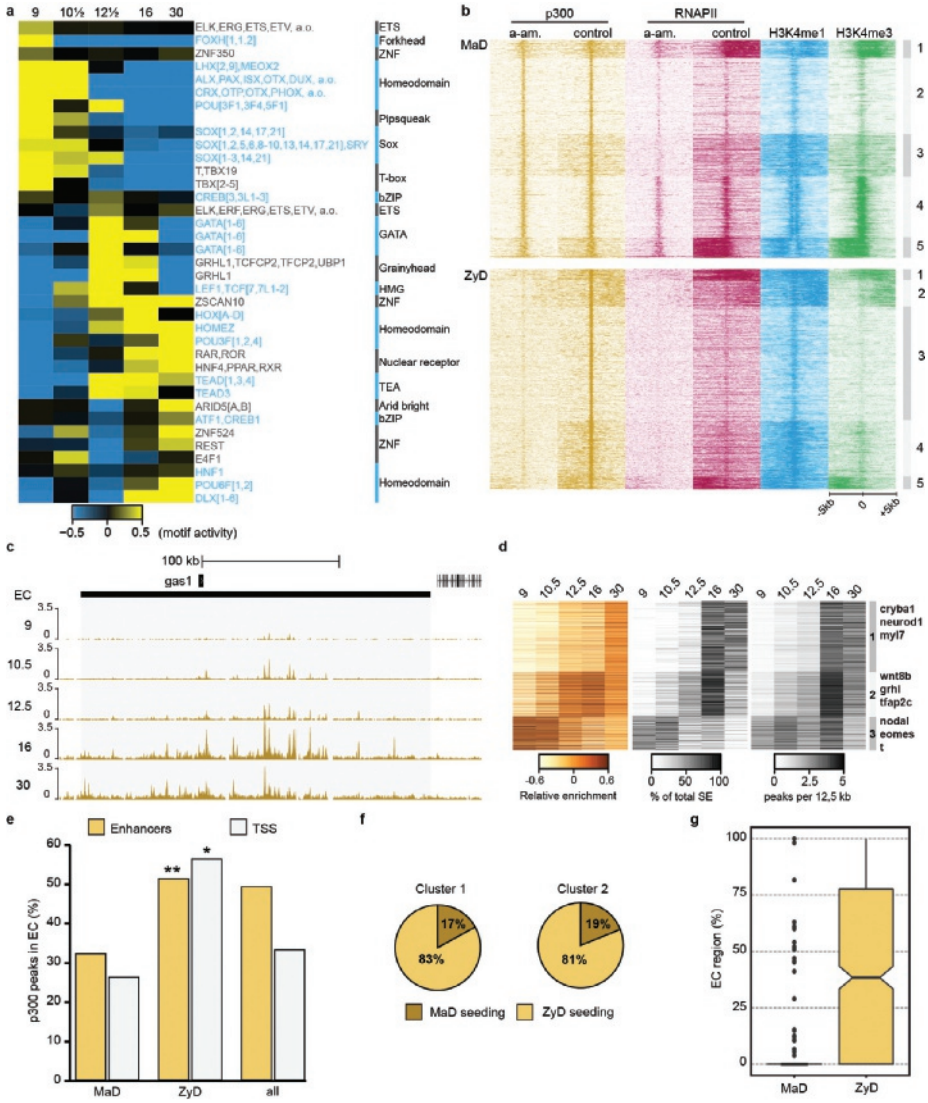


Figure 5. Zygotically controlled p300 recruitment shapes enhancer clusters (EC) domains. **(a)** Modelled transcription factor motif activity to p300 enrichment (see Methods). Activity reflects modelled contributions in p300 peak RPKM. **(b)** Heatmaps of MaD (upper panel) and ZyD (lower panel) p300 binding sites in a-amanitin treated and control embryos. **(c)** Developmental increase in genomic coverage of the gas1 EC by acquisition of p300 binding at enhancers. **(d)** EC dynamics of p300 enrichment (left panel), percentage of total EC region identified in each stage based on stage-dependent p300 binding (middle panel) and number of p300 peaks (per 12.5kb) in EC. **(e)** Percentage of zygotic defined (ZyD, lost in a-amanitin treated embryos) and maternal defined (MaD, maintained in a-amanitin treated embryos) p300 peaks that map to ECs. Asterisks indicate significance as more or less p300 peaks than expected by chance calculated using cumulative hypergeometric test: *P=6E-14; **P=5E-29 **(f)** Percentage of ECs that have a MaD or ZyD seeding peak at stage 9. **(g)** Box plot showing the percentage of the EC region that

is defined by MaD or ZyD p300 peaks. Box: 25th (bottom), 50th (internal band), 75th (top) percentiles. Whiskers: 1.5×interquartile range of the lower and upper quartiles, respectively. Outliers are indicated with black dots.

ZyD p300-bound domains shape enhancer clusters

P300 can be recruited by transcription factors that bind to regulatory elements. We therefore modelled transcription factor motif contributions to p300 binding across multiple developmental stages (see Methods). The results predict specific transcription factors to recruit p300 in a stage-specific manner (Fig. 5a). Clustering of MaD and ZyD p300-bound regions with H3K4me3, H3K4me1 and RNAPII data revealed that ZyD p300 is recruited to distal regulatory sequences that lose both p300 and RNAPII binding in the presence of α-amanitin, whereas MaD p300 binding mostly includes promoter-proximal regions that are H3K4me3-decorated and recruit RNAPII in the presence of α-amanitin but without elongating (Fig. 5b). Indeed, MaD p300 regions are enriched for promoter-related motifs (Supplementary Fig. 7). Although some ZyD p300-bound regions overlap with annotated transcription start sites (Supplementary Fig. 5f), most of these sequences are decorated with H3K4me1 in the absence of H3K4me3, suggesting they correspond to distal regulatory sequences (Fig. 5b). Both MaD and ZyD p300-bound regulatory regions recruit embryonically regulated transcription factors such as Otx2, Gsc, Smad2/3, Foxh1, T (Xbra), Vegt and Eomes (Supplementary Fig. 8)[39-41], suggesting that multiple transcription factors contribute to p300 recruitment.

Large enhancer clusters (ECs) are thought to improve the stability of enhancer-promoter interactions, are associated with genes coding for developmental regulators, and have been implicated in cell differentiation[42-44]. During development the cluster size of p300-bound enhancers grows dynamically by p300 seeding of individual enhancers (Fig. 5c, d, see Methods). Histone modifications and transcript levels of EC-associated genes are developmental stage specific, confirming the association of ECs with developmental genes (Supplementary Fig. 9; Supplementary Data 2). Analysis of the percentage of the total EC regions identified in each stage show that most p300-bound ECs increase in genomic coverage during development by newly gained p300 binding at enhancers (EC clusters 1 and 2), whereas a group of early ECs (EC cluster 3) decrease in coverage as a result of the decreasing number of p300 peaks that contribute to the EC.

We next examined how MaD and ZyD p300-bound regions contribute to p300-bound ECs. Approximately 50% of all ZyD p300-bound enhancers are located in ECs at stage 11. Among MaD p300-bound enhancers this fraction is much reduced (Fig. 5e). Similarly, a much larger fraction of ZyD p300-bound promoters is found in ECs compared with MaD p300-bound promoters. Up to 20% of the developmental ECs that are seeded at stage 9 have a MaD p300

seeding site (Fig. 5f). However, very few ECs can be called based on MaD p300, showing that formation of p300-bound enhancer clusters requires embryonic transcription (Fig. 5g).

Extended maternal epigenetic control

We next examined the extent to which the MaD epigenome is maintained during development. Genes were grouped based on MaD or ZyD trimethylation of H3K4 and H3K27 in the promoter (Supplementary Data 3, see Methods). For p300 we counted the total number of MaD and ZyD peaks in the cis-regulatory landscapes of genes (Fig. 6a). Remarkably, MaD H3K4me3-regulated genes represent the majority of all H3K4me3-enriched genes in both early and late developmental stages. Even at neurula and tailbud stages only a small fraction of the H3K4me3-decorated genes are ZyD. Similarly, maternal control of H3K27me3 also extends late into development, albeit to a smaller degree. After gastrulation, the number of MaD H3K27me3 regulated genes slightly decreases, whereas ZyD increases. However, also at neurula stage more than 50% of the Polycomb (PRC2)-regulated genes are under MaD H3K27me3 control. By contrast, p300 in cis-regulatory regions of genes is almost exclusively ZyD in all stages (Fig. 6a).

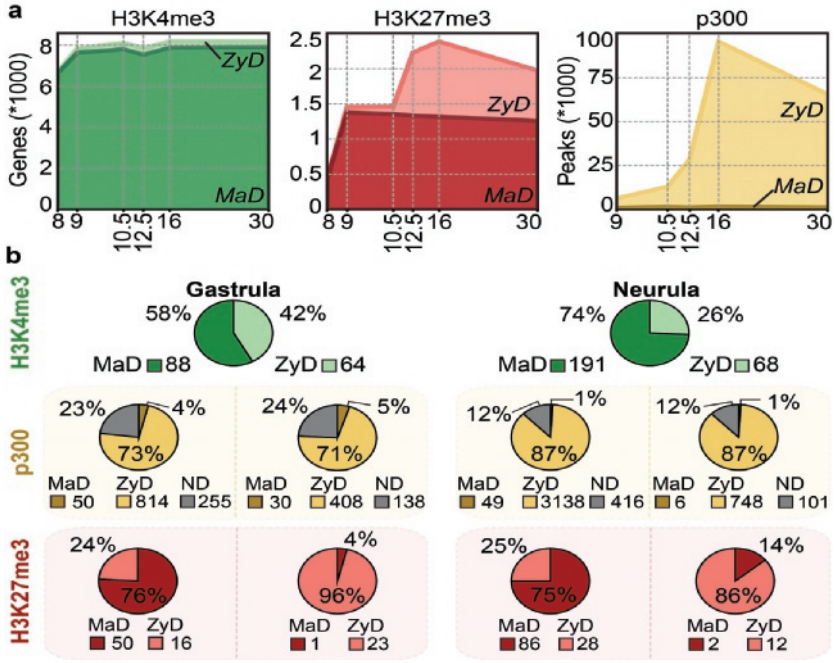


Figure 6. Maternal epigenetic control extends beyond gastrulation. Maternally defined (MaD) peaks emerge at or before stage 11 independent of embryonic transcription. Zygotically defined (ZyD) peaks appear before stage 11 and are lost in a-amanitin treated embryos, or emerge at or after stage 12. Not determined (ND) peaks are not consistently detected in replicates 1 and 2 and generally have low enrichment values. **(a)** Total number of genes with a MaD or ZyD peak in their promoter (H3K4me3 and H3K27me3), or total number of MaD and ZyD peaks per GREAT region (p300). ND peaks are not shown. **(b)** MaD and ZyD regulation of gastrula and neurula expressed genes. The pie charts show the number genes with a MaD or ZyD peak in their promoter (H3K4me3 and H3K27me3) or the number of MaD, ZyD and ND peaks per cis-regulatory region (p300). The H3K27me3 and p300 pie charts represent: Gastrula expressed genes with a MaD (far left) or ZyD (middle left) H3K4me3 peak; neurula expressed genes with a MaD (middle right) or ZyD (far right) H3K4me3 peak.

Many genes may maintain MaD H3K4me3 because they are constitutively expressed throughout development. We therefore analysed the regulation of genes that are exclusively embryonically transcribed. We find that 487 of 983 (49.5%) genes which are expressed between blastula and tailbud stages but not expressed in oocytes or before the MBT, feature a MaD H3K4me3 promoter (Supplementary Fig. 10a). Most of the MaD H3K4me3 genes that are modified by PRC2 exhibit MaD H3K27me3. When separating embryonic transcripts based on developmental activation, we find MaD H3K4me3 for 58% of the gastrula genes and up to 74% of the neurula expressed genes (Fig. 6b; Supplementary Fig. 10b). In most cases MaD H3K4me3-

regulated genes also have MaD H3K27me3 control. This indicates an important role for the MaD epigenome in the regulation of embryonic transcripts.

To explore the distinctions between expression inside and outside the maternal regulatory space, we analysed Wnt signalling targets. Early Wnt/beta-catenin signalling serves to specify dorsal fates following fertilization, leading to organizer gene expression. This has been shown to depend on Prmt2-mediated promoter poising before the MBT[45]. Indeed, we find that seven of eight early Wnt/beta-catenin targets have a hypomethylated island promoter marked with MaD H3K4me3 (Fig. 7a; Supplementary Fig. 10c). Wnt signalling also plays an important role after the MBT, when it ventralises and patterns mesoderm. The majority of these later targets turn out to have a methylated promoter with ZyD H3K4me3. Notably, these ZyD H3K4me3 late Wnt targets are associated with high binding of p300 in their locus; many of the p300 binding events happen at distal regulatory regions. In contrast, MaD H3K4me3 Wnt targets have less p300 binding but are marked with H3K27me3 (Fig. 7a, b). These results illustrate the dichotomy in proximal and distal regulation that is associated with transcriptional activation of maternal and zygotic Wnt target genes, which is paradigmatic of the distinctive maternal and zygotic epigenetic programs that are orchestrated by DNA methylation and exert a long-lasting influence in development (Fig. 8).

DISCUSSION

The H3K4me3 modification poises promoters for transcription initiation by stabilizing Taf3/TFIID binding 36, 37. Promoter H3K4 methylation based on an underlying DNA methylation logic driven by maternal factors at the blastula stage sets the stage for a default program of gene expression. Most constitutively expressed house-keeping genes are within this maternal regulatory space, as well a subset of developmentally regulated genes. Remarkably, many late expressed genes have hypomethylated promoters and are already poised for activation by H3K4me3 during early blastula stages. H3K4me3 is not sufficient for gene transcription and additional embryonic factors are required for activation in many cases. Genes with MaD H3K4me3 generally have fewer p300-bound enhancers associated with them, suggesting they are regulated by promoter-proximal elements. This further underscores the permissive nature of this regulation, as opposed to zygotically regulated events at both promoters (H3K4me3) and enhancers (recruitment of p300). The H3K27me3 modification is gradually acquired between blastula and gastrula stages on spatially regulated genes, repressing lineage-specific genes in other lineages[13, 14]. The acquisition of this modification in the absence of transcription indicates that it is uncoupled from the inductive events of the early embryo, suggesting a default maternal response to a lack of transcriptional activation. The results indicate that maternal factors set permissions and time-dependent constraints on a subset of genes with

reduced CpG methylation at their promoter. These permissions and constraints are likely to channel embryonic cell fates into a limited number of directions by controlling hierarchical developmental progression by master regulators. Previously we observed that DNA methylation does not lead to transcriptional repression in early embryos, whereas it does in oocytes and late embryos[46]. The observations described here suggest a new role of DNA methylation in defining a maternal-embryonic program of gene expression. In zebrafish, the maternal methylome is reprogrammed between fertilization and ZGA, to match the paternal methylome. This also occurs in maternal-haploid fish, and appears to align with CG content[47, 48], suggesting an intrinsic maternal mechanism that sets the stage for the MaD epigenome.

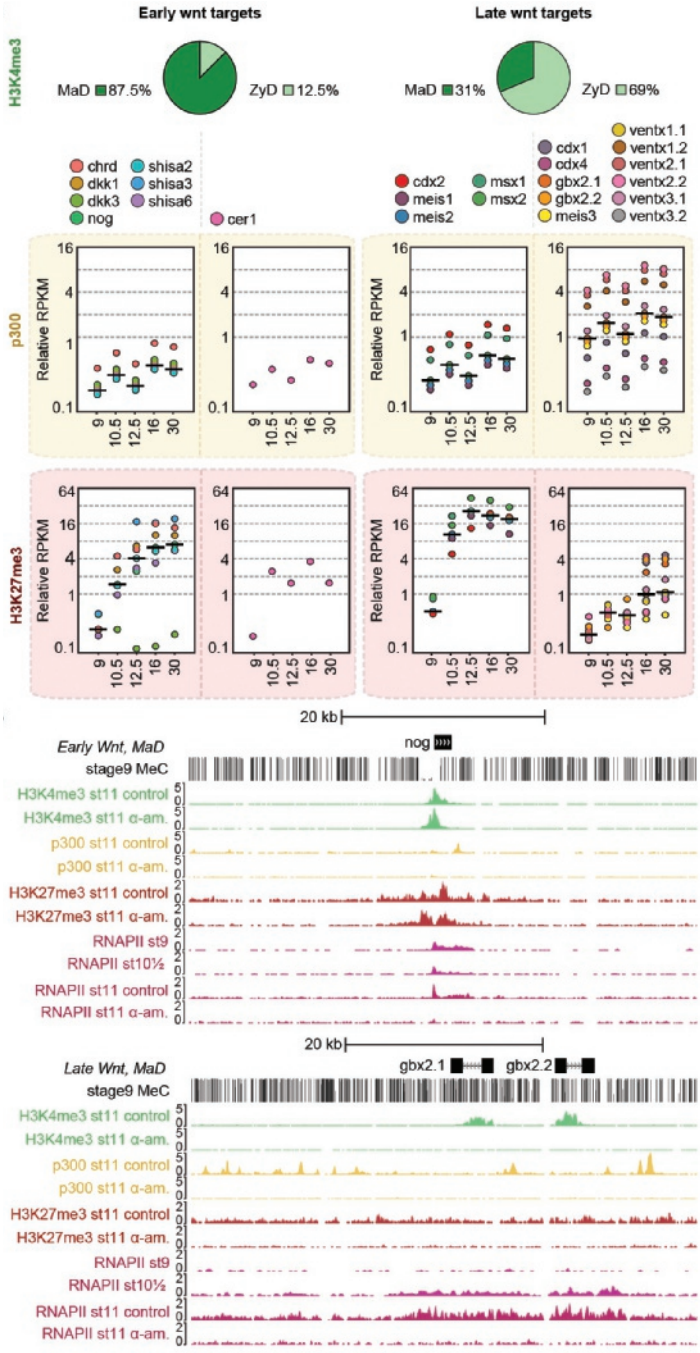


Figure 7. Maternal and zygotic regulatory space separates early and late Wnt target genes. **(a)** The number of genes with MaD or ZyD H3K4me3 (pie charts) and relative RPKM (dot plots, horizontal line: median) of p300 in cis-regulatory regions of genes and H3K27me3 on promoters (±2.5 kb from TSS)

at different developmental stages that have maternally or zygotically defined H3K4me3 at the promoter. Early targets sia1 and sia2 are not included, these genes lose H3K4me3 after stage 9 and cannot be assigned to MaD or ZyD space based on our stage 11 a-amanitin data. H3K4me3 on these genes is acquired at stage 8, before embryonic transcription. **(b)** Browser views of the early Wnt target nog (noggin) and the late Wnt targets gbx2.1 and gbx2.2 with ChIP-seq enrichment of H3K4me3, p300 and RNAPII on control and a-amanitin-injected embryos and RNAPII on stages 9 and 10.5.

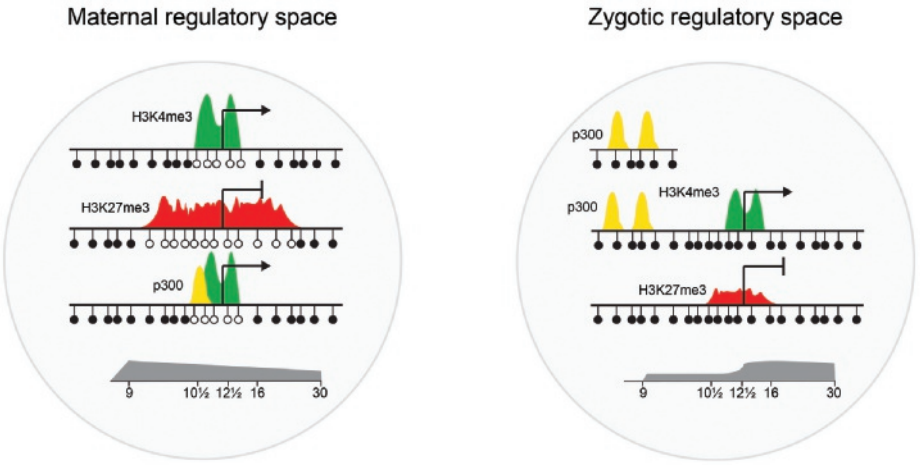


Figure 8. Model of maternal and zygotic regulatory space. This shows the segregation of maternal regulatory space, which contains hypomethylated promoters that are mainly controlled by maternal factors, and zygotic regulatory space, which includes methylated promoters and enhancers that are under zygotic control. Most p300-bound enhancers are in zygotic space, however, they can regulate promoters in both maternal and zygotic space, crossing the regulatory space border. This may contribute to varying degrees of permissiveness to transcriptional activation. Maternal regulatory space extends well into neurula and tailbud stages and includes many embryonic genes which are activated at specific stages of development. Zygotic regulatory space requires zygotic transcription, is established from the mid-blastula stage onwards but increases in relative contribution during development.

Gene expression outside maternal regulatory space could be mediated by p300-associated enhancers, most of which require new transcription for recruitment of p300. Promoter and enhancer activation in the ZyD regulatory space likely involves binding of specific factors. Indeed, we find that both MaD and ZyD p300-bound regulatory regions recruit embryonically regulated transcription factors. Enhancers often contain binding sites for many different proteins, which can play different roles in opening up chromatin, recruitment of co-activators and establishing looping interactions with promoters. Future experiments will shed light on the maternal-zygotic hierarchy and the regulatory transitions underlying these events and the roles of maternal and zygotic pioneer factors. We find that ZyD p300-bound enhancers shape enhancer clusters. These form dense hubs of regulatory activity, and EC p300 binding

is generally correlated with the expression of the associated genes. The work reported here suggests that recruitment of p300 to "seeding" enhancers precedes establishing cluster-wide activity of the local enhancer landscape. Future work will also need to address to which extent seeding causes relaxation and opening of the local chromatin and activity of neighbouring enhancers.

Key proteins of the molecular machinery involved in DNA methylation (Dnmt3a, Tet2), H3K4me3 (Mll1-4, Kdm5b/c), H3K27me3 (Ezh2, Eed, Kdm6a/b) and enhancer histone acetylation (p300) are not only highly conserved between species but also frequently mutated in cancer 49, 50, 51. Moreover cancer-specific hypermethylated regions tend to correspond to Polycomb-regulated loci in embryonic stem cells and DNA methylation may restrict H3K27 methylation globally 52, 53. In addition, the sequence signatures of hypomethylated regions that acquire H3K4me3 or H3K27me3 are conserved between fish, frogs and humans 13. These observations suggest that the molecular mechanisms that orchestrate the maternal and zygotic regulatory space are conserved. One key difference between mammals and non-mammalian vertebrates is the specification of extra-embryonic lineages between zygotic genome activation and the blastocyst stage in mammals 10, so it is likely that the way this plays out for specific genes differs between species. In summary, our results provide an unprecedented view of the far reach of maternal factors in zygotic life through chromatin state. The dichotomy of maternal promoter-based and embryonic enhancer regulation demarcates an epigenetic maternal-to-zygotic transition that is maternal-permissive to the expression of some embryonic genes and restrictive to others. This highlights the combinatorial interplay of maternal and zygotic factors through distinct mechanisms.

METHODS

Animal procedures

X. tropicalis embryos were obtained by in vitro fertilization, dejellied in 3% cysteine and collected at the indicated stage. Fertilized eggs were injected with 2.3 nl of 2.67 ng/µl α-amanitin and developed until the control embryos reached mid-gastrulation (stage 11). Animal use was conducted under the DEC permission (Dutch Animal Experimentation Committee) RU-DEC 2012–116 and 2014–122 to G.J.C.V..

ChIP-sequencing and RNA-sequencing

Chromatin for chromatin-immunoprecipitation (ChIP) was prepared as previously described[54, 55], with minor modifications. Antibody was incubated with chromatin overnight, followed by incubation with Dynabeads Protein G for 1 h. The following antibodies were used: anti-H3K4me1 (Abcam ab8895, 1 µg per 15 embryo equivalents (Eeq)), anti-H3K4me3 (Abcam ab8580, 1 µg per 15 Eeq), anti-H3K9ac (Upstate/Millipore 06-942, 1 µg per 15 Eeq), anti-H3K36me3 (Abcam ab9050, 1 µg per 15 Eeq), anti-H3K27me3 (Upstate/Millipore 07-449, 1 µg per 15 Eeq), anti-H3K9me2 (Diagenode C15410060, 1 µg per 15 Eeq), anti-H3K9me3 (Abcam ab8898, 2 µg per 15 Eeq), anti-H4K20me3 (Abcam ab9053, 2 µg per 15 Eeq), anti-p300 (Santa Cruz sc-585, 1 µg per 15 Eeq), and anti-RNAPII (Diagenode C15200004, 1 µg per 15 Eeq). For all ChIP-seq samples of the epigenome reference maps and RNAPII ChIP-seq samples of the α-amanitin experiments three biological replicates of different chromatin isolations of 45 embryos were pooled. Two biological replicates for H3K4me3 (α-amanitin injected: resp. 90 and 56 embryo equivalents (Eeq); control: resp. 45 and 67 eeq), H3K27me3 (α-amanitin injected: resp. 90 and 180 Eeq; control: resp. 45 and 202 eeq) and p300 (α-amanitin injected: resp. 112 and 56 Eeq; control: resp. 112 and 67 Eeq) ChIP-seq samples of the α-amanitin experiments were generated. For RNA-seq samples of the α-amanitin experiments RNA from five embryos from one biological replicate was isolated and depleted of ribosomal RNA as previously described[35]. Samples were subjected to a qPCR quality check pre- and post-preparation. Libraries were prepared with the Kapa Hyper Prep kit (Kapa Biosystems), and sequencing was done on the Illumina HiSeq2000 platform. Reads were mapped to the reference *X. tropicalis* genome JGI7.1, using STAR (RNA-seq) or BWA (ChIP-seq) allowing one mismatch.

MethylC-seq

Genomic DNA from *Xenopus* embryos stages 9 and 10.5 was obtained as described before[56]. MethylC-seq library generation was performed as described previously[57]. Library amplification was performed with KAPA HiFi HotStart Uracil+ DNA polymerase (Kapa Biosystems, Woburn, MA, USA), using six cycles of amplification. Single-read MethylC-seq libraries were processed and aligned as described previously[58].

Quantitative PCR (qPCR)

PCR reactions were performed on a CFX96 Touch Real-Time PCR Detection System (BioRad) using iQ Custom SYBR Green Supermix (BioRad). We preformed RNA expression PCR (RT–qPCR (quantitative PCR)) and ChIP-qPCR for H3K4me3 and H3K9ac on promoters of odc1, *eef1a1o*, rnf146, tor1a, zic1, cdc14b, eomes, xrcc1, drosha, gdf3, t, tbx2, fastkd3, *gs17* (see Supplementary Methods for primer sequences). ChIP-qPCR enrichment over background was calculated using the average of 5 negative loci.

Detection of enriched regions

We used MACS2[59] with standard settings and a q-value of 0.05. Fragment size was determined using phantom-peakqualtools[60]. Broad settings (--BROAD) were used for H3K4me1, H3K36me3, H3K27me3, H3K9me2, H3K9me3, H4K20me3 and RNAPII. Broad and narrow peaks were merged for H3K4me3. For H3K9ac narrow peaks were used. For p300 broad peaks were used in the ChomHMM analysis, narrow p300 peaks were used for super-enhancer and MaD versus ZyD analyses. All peaks were called relative to an input control track. Peaks that showed at least 75% overlap with 1 kb regions that have more than 65 input reads, and peaks that have a ChIP-seq RPKM higher than the 95 percentile of random background regions are excluded from further analysis. Only scaffolds 1-10 (the chromosome-sized scaffolds) were included in the analysis. Relative RPKM was calculated by dividing the ChIP-seq RPKM of the peaks by the ChIP-seq RPKM of the 95 percentile of random background regions.

We used MAnorm[61] to determine differentially enriched regions in α-amanitin and control embryos. We used merged peak sets of replicate 1, replicate 2 and stage 10.5 to avoid bias caused by peak calling. Lost, gained and unchanged peaks per biological replicate were determined using the following parameters: lost peaks have M-values > 1 and a -log base 10(P value) > 5 (for H3K27me3) or 1.3 (for H3K4me3 and p300) and have a relative RPKM (background corrected) > 1 in stage 11 control (no cut-off was used for st.11 control of H3K27me3 rep.1), stage 10.5 (H3K4me3 and p300) or stage 12 (H3K27me3); increased peaks have M-values smaller than -1 and a -log base 10(p-value) > 5 (H3K27me3) or 1.3 (H3K4me3 and p300) and have a rel. RPKM greater than 1 in stage 11 α-amanitin, stage 10.5 (H3K4me3 and p300) or stage 12 (H3K27me3); unchanged peaks are neither gained nor lost and have a relative RPKM > 1 in stage 11 control (no cut-off was used for st.11 control of H3K27me3 rep.1), stage 11 α-amanitin, stage 10.5 (H3K4me3 and p300) or stage 12 (H3K27me3). Maintained peaks are peaks that are not lost and have a rel. RPKM greater than 1 in stage 11 control (no cut-off was used for st.11 control of H3K27me3 rep.1), stage 11 α-amanitin, stage 10.5 (H3K4me3 and p300) or stage 12 (H3K27me3). Common lost, gained, unbiased and maintained peaks are present in both replicates. All other peaks are considered not defined (ND). Replicate-specific peaks were only used for Supplementary Fig. 4b, for all other figures the common peaks were used.

DNA methylation levels in Supplementary Fig. 4d was calculated using previously published Bio-CAP data[62]. Bio-CAP RPKM levels of stage 11-12 were calculated for H3K4me3, H3K27me3 and p300 peaks, and corrected for input values. For Fig. 4c genes were considered "hypomethylated" if the Bio-CAP/Input ratio on the promoter (± 1 kb from TSS) was > 1.

RNA expression analysis was performed as previously published[35]. Embryonic transcripts were separated based on the clustering of maximum expression levels per stage in Fig. 3d of Paranjpe et al.[35] (cluster 1 = blastula, cluster 5 = gastrula, clusters 3 and 4 = neurula, clusters 2 and 6 = tailbud).

Enhancer clusters were called as previously described[43]. Enhancer Clusters are called per stage and merged to determine the total Enhancer Cluster region. Percentage of the EC region is calculated relative to the total Enhancer Cluster region.

MaD and ZyD classification

Maternally defined (MaD) peaks emerge at or before stage 11 and are also acquired in α-amanitin treated embryos in both replicates. Zygotically defined (ZyD) peaks appear at or before stage 11 and are lost in α-amanitin treated embryos in both replicates, or emerge after stage 11. To classify MaD and ZyD H3K4me3 genes we ran MAnorm on promoters (+ 250 bp from TSS) only, using similar restrictions as described in Detection of enriched regions. MaD H3K4me3 genes have a maintained promoter in both replicates, ZyD H3K4me3 genes have a lost promoter H3K4me3 peak in both α-amanitin replicates, or a peak that emerges after stage 11. MaD H3K27me3 genes have at least one MaD peak in the vicinity of their promoter (+ 2.5 kb from TSS). ZyD H3K27me3 genes have at least one ZyD peak in their promoter and lack a MaD peak. Not defined (ND) peaks or genes do meet the criteria for neither MaD nor ZyD. For p300 the total number of ZyD and MaD peaks was counted in GREAT63 regions of genes.

ChomHMM analysis

Chromatin states were discovered and characterized using ChromHMM v1.1026, an implementation of a hidden Markov model. As input we used the enriched regions from ten tracks (H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me2, H3K9me3, H4K20me3, p300 and RNAPII) across five developmental stages. We trained and ran the model with a range of states, and determined the 19 emission states model as the optimal number of states that could sufficiently capture the biological variation in co-occurrence of chromatin marks. We subsequently classified the states into 7 main groups based on the presence and absence of specific chromatin marks.

The segmentation files of the 7 main groups per stage were binned in 200 base pairs intervals. An m × n matrix was created, where m corresponds to the 200 base-pair intervals and n to the developmental stages (9-30). Each element a(i,j) represents the chromatin state of interval i at stage j. For each chromatin group occurrences were counted per stage n. The changes between stage n and n+1 were plotted using Sankey diagrams (https://github.com/tamc/Sankey), a flow diagram closely related to alluvial diagrams.

Motif analyses

For the prediction of motif contribution to p300 recruitment (Fig. 5a) we have implemented the ISMARA method developed by Balwierz et al.64. This method uses motif activity response analysis to determine the transcription factors that drive the observed changes in chromatin state across samples. As input we used the number of known motifs found per p300 binding site and the RPKM of the p300 peaks per developmental stage. The model infers the unknown motif activities from the equation in which the changes in signal levels are explained with the number of binding sites and the unknown motif activities. Motifs that showed a z-score activity that was higher than 13 are shown in Fig. 5a. Enriched motifs (Supplementary Fig. 7) were detected with gimme diff, a tool from the GimmeMotifs package 65. The vertebrate motifs used in this script were obtained from CIS-BP (http://cisbp.ccbr.utoronto.ca/)66 and clustered using gimme cluster from GimmeMotifs. The motifs are available at http://dx.doi.org/10.6084/m9.figshare.1555851 (Van Heeringen, Simon J. (2015): Vertebrate motif clusters v3.0. figshare.).

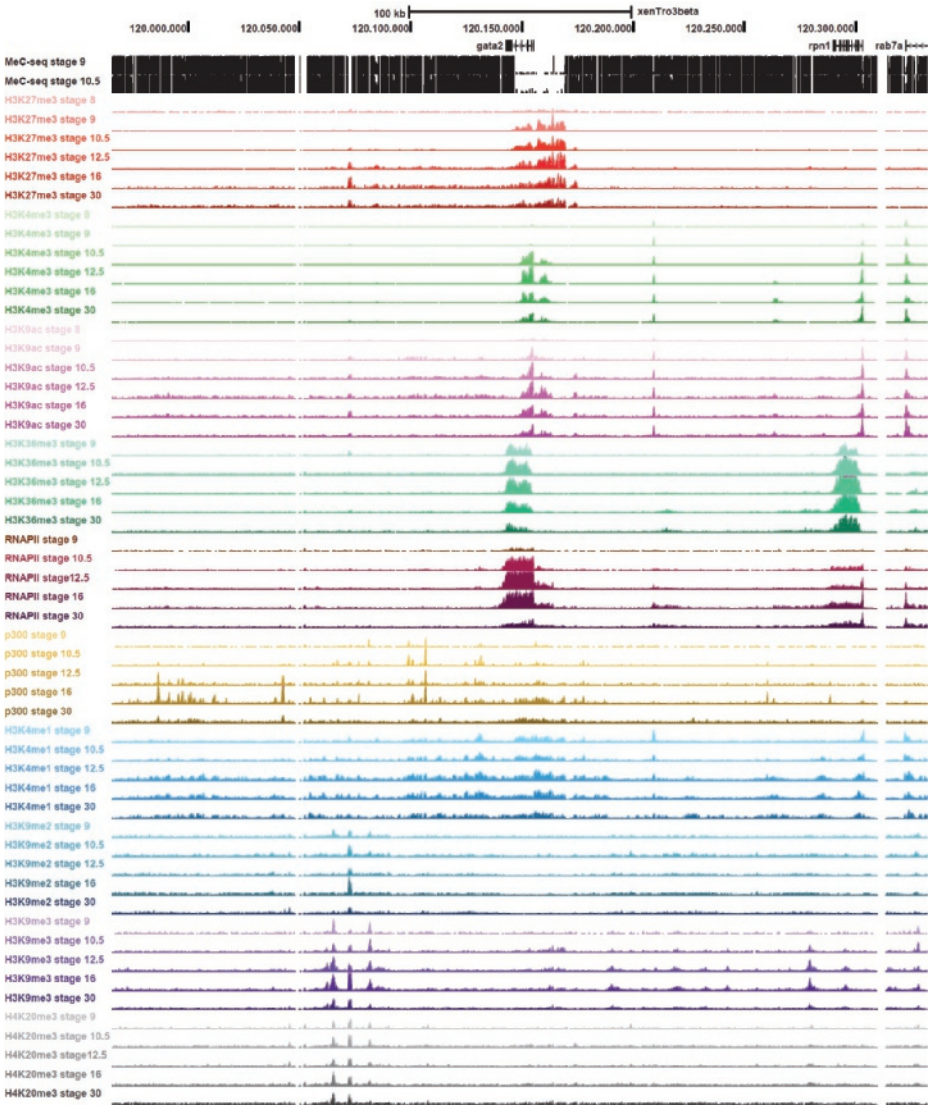
Generation of plots and heatmaps

All heatmaps were generated using fluff (http://simonvh.github.com/fluff)13 or gplots (http://cran.r-project.org/web/packages/gplots/index.html). For all heatmap clustering, the Euclidean distance metric was used. Other plots were generated using ggplot2 (http://ggplot2.org/).

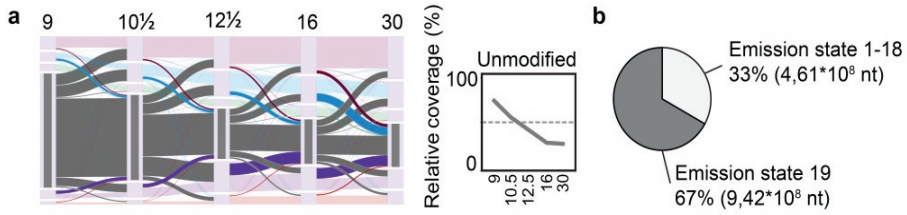
Data accessibility

The data generated for this work have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE67974. Visualization tracks are available at the authors' web site (http://www.ncmls.nl/gertjanveenstra).

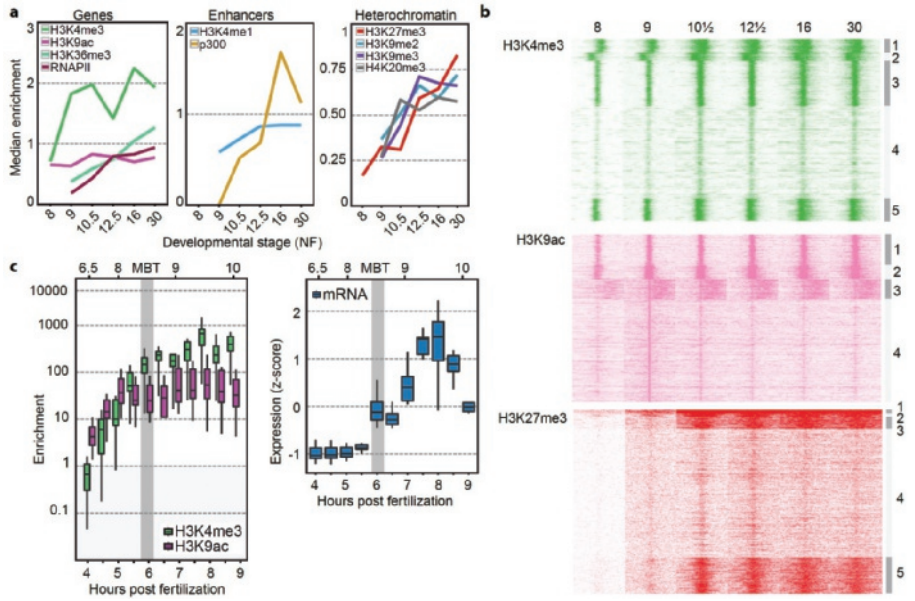
SUPPLEMENTARY FIGURES



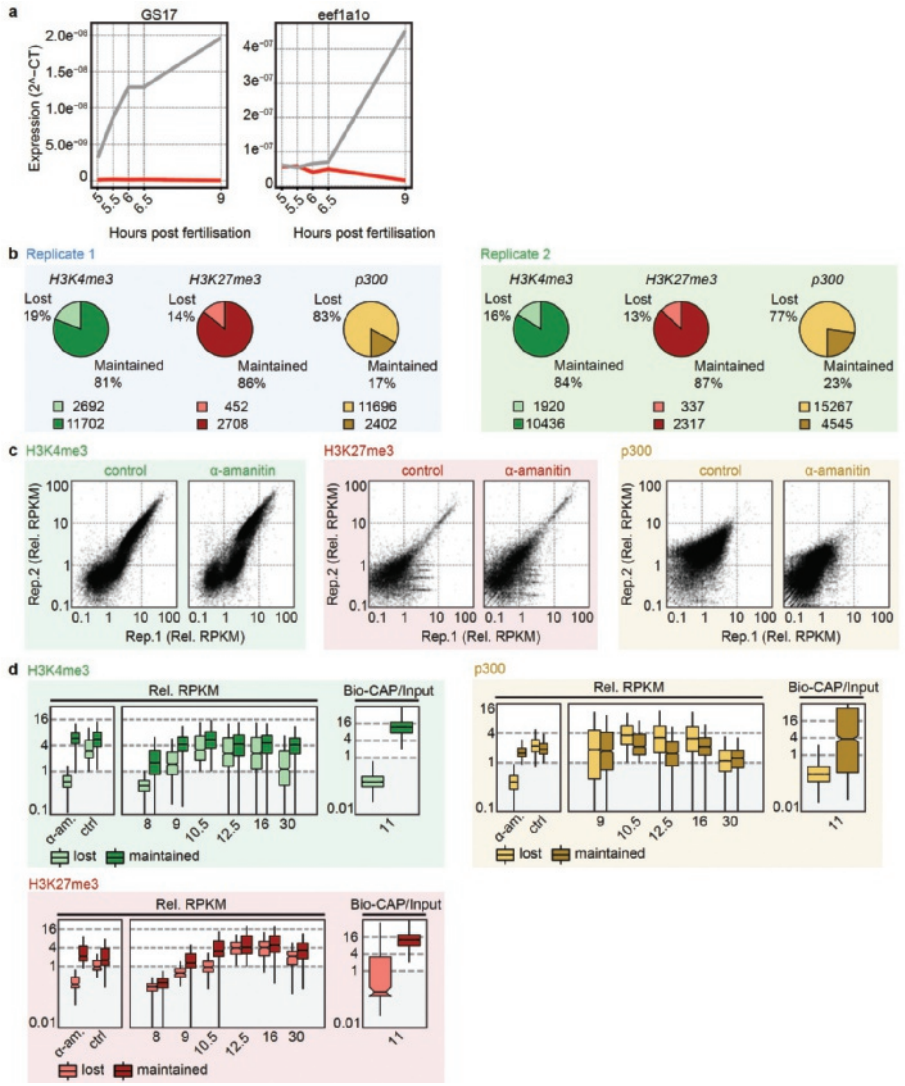
Supplementary Figure 1. Gata2 locus. ChIP-seq enrichment of histone modifications, RNAPII and p300 for stages 9- 30. The heterochromatin tracks (H3K9me2, H3K9me3 and H4K20me3) are shown including non-unique sequence reads, identifying repetitive regions enriched for these modifications. ChIP-sequencing on stage 8 (blastula, pre-MBT) was done for histone modifications H3K4me3, H3K9ac and H3K27me3.



Supplementary Figure 2. Unmodified state coverage. (a) Alluvial plots of unmodified state (grey) coverage during development. The height represents the fraction of the modified genome that contributes to the same or a different chromatin state. The line plots shows coverage of the unmodified state per stage as a percentage of the sum of all regions that are state 1-18 at any stage. **(b)** Absolute nucleotide coverage of emission state 19 and states 1-18 at stage 30. It should be noted that 'unmodified' specifically refers to the examined histone modifications and that this state shows abundant DNA methylation.

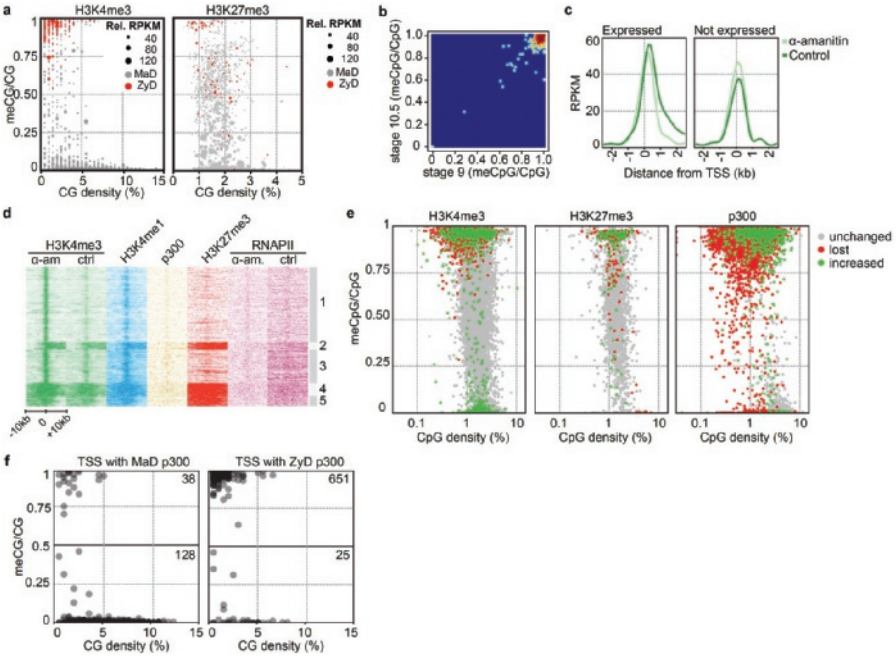


Supplementary Figure 3. Progressive specification of the epigenome. (a) Median enrichment of chromatin marks during development. **(b)** RPKM levels of H3K4me3, H3K9ac and H3K27me3 stage 8-30 on stage 9 peaks. Most stage 9 H3K4me3 and H3K9ac peaks show already significant enrichment at stage 8, whereas H3K27me3 markedly increases in late blastula and early gastrula embryos. **(c)** Detailed time series from 4 to 9 hours post fertilization (13 genes, average values of two biological replicates, see Methods). Left panel: Box plot of ChIP-qPCR for H3K9ac (pink) and H3K4me3 (green). Right panel: Box plot of RNA expression (RT-qPCR). Box: 25th (bottom), 50th (internal band), 75th (top) percentiles. Whiskers: 1.5 * interquartile range of the lower and upper quartiles, respectively.

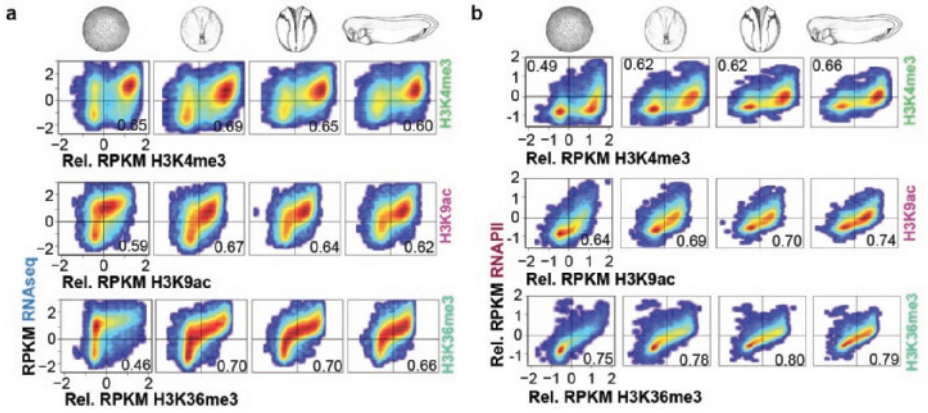


Supplementary Figure 4. Maternal and zygotic acquisition of chromatin state. (a) RNA expression (RT-qPCR) of *gs17* (embryonic transcript), *eef1a1o* (maternal transcript, induced at MBT) in α-amanitin and control embros. **(b)** Lost and Maintained peaks of H3K4me3, H3K27me3 and p300 in replicate 1 (left, blue background) and replicate 2 (right, green background). Pie charts representing percentage and numbers of lost and maintained peaks per replicate. **(c)** Scatter plots with relative RPKM (background corrected) of replicate 1 (x-axis) and replicate 2 (y-axis) on peaks that are lost or maintained in both experiments. **(d)** Left and middle panels show box plots of relative RPKM (background corrected) of regions with MaD or ZyD H3K4me3, H3K27me3 or p300-binding.

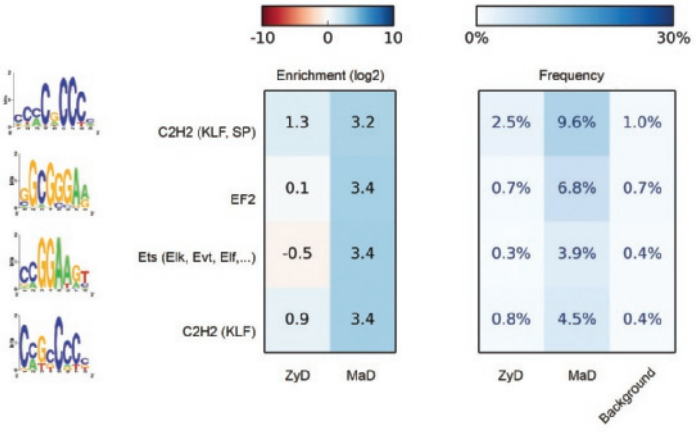
Right panels show box plots of input corrected RPKM of previously profiled Bio-CAP data representing hypomethylated DNA domains [61]. MaD trimethylation of H3K4 and H3K27 is detected almost exclusively on Bio-CAP-enriched regions indicating clusters of hypomethylated CpGs. Box: 25th (bottom), 50th (internal band), 75th (top) percentiles. Whiskers: 1.5 * interquartile range of the lower and upper quartiles, respectively.



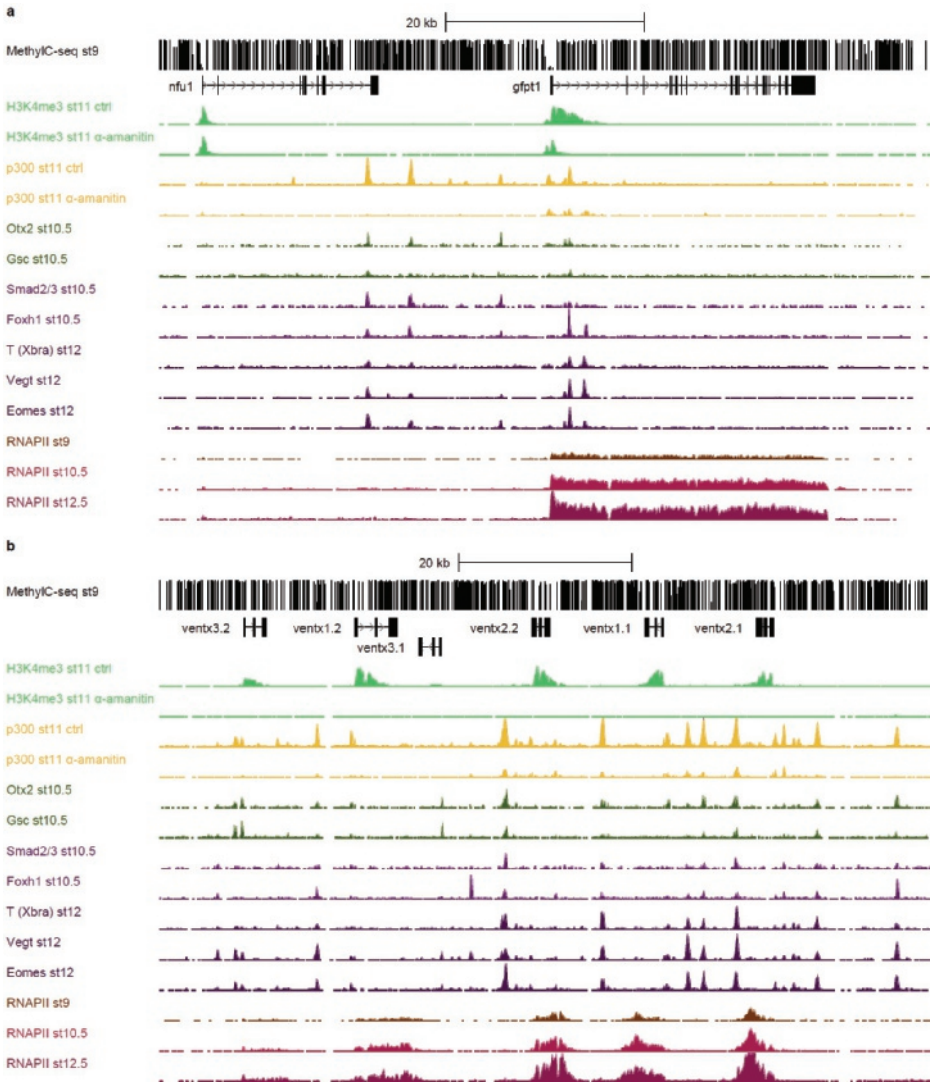
Supplementary Figure 5. Methylation logic for maternal and zygotic defined chromatin state. (a) CpG density and methylation at stage 10.5 of promoters (H3K4me3: ± 100 bp from TSS; H3K27me3: ± 2.5 kb from TSS) that contain a zygotic defined (ZyD, lost in α-amanitin treated embryos, red) or maternal defined (MaD, maintained in α-amanitin treated embryos, grey) peak for H3K4me3 (left) or H3K27me3 (right) after inhibition of embryonic transcription. The size of the dot indicates the relative RPKM (background corrected). **(b)** Density heatmap of DNA methylation stage 9 (x-axis) and stage 10.5 (y-axis) on ZyD promoters (± 100 bp from TSS). **(c)** Mean relative RPKM of stage 11 α-amanitin and control H3K4me3 on promoters of stage 10.5 expressed (left) and not expressed genes (right). **(d)** Heatmap representation of regions with increased H3K4me3 deposition in α-amanitin treated embryos**. (e)** CG density and methylation (stage 9) on lost, increased and unchanged H3K4me3 (left), H3K27me3 (middle) or p300 (right) peaks. For the purpose of simplicity, unchanged and increased peaks are collectively referred to as MaD in the rest of this article. **(f)** CG density and methylation on promoters (± 100 bp from TSS) that overlap with MaD (left) or ZyD (right) p300-bound peaks. The values in the middle and top corners indicate the number of promoters with meCG/CG ratio above or below 0.5.



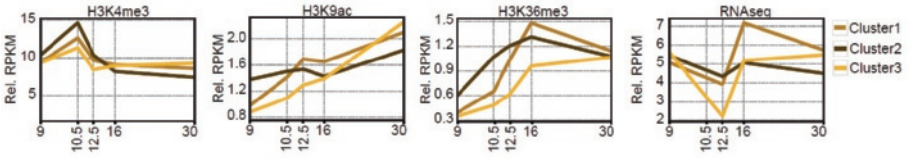
Supplementary Figure 6. Correlation of chromatin marks and transcription. Density correlation plots of relative RPKM (background corrected) for H3K4me3 and H3K9ac (± 1 kb from TSS) and H3K36me3 (genes bodies) with **(a)** RNAseq (exons) or **(b)** RNA polymerase II (gene bodies).



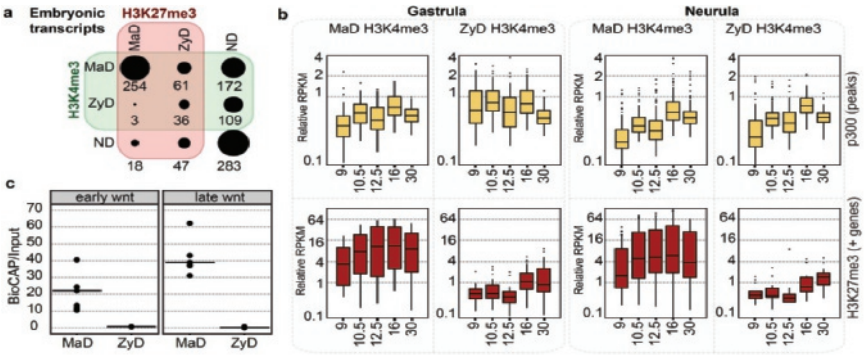
Supplementary Figure 7. MaD p300 regions are enriched for promoter related motif sequences. Motif enrichment and frequency in MaD and ZyD p300-bound regions.



Supplementary Figure 8. MaD and ZyD p300 bound regions recruit embryonically regulated transcription factors. *Gfpt1* **(a)** and *ventx* **(b)** locus with stage 9 MethylC-seq and ChIP-seq enrichment of H3K4me3 and p300 on control and α-amanitin injected embryos, transcription factors Otx2, Gsc, Smad2/3, Foxh1, T (Xbra), Vegt, Eomes and RNAPII on stage 9 10.5 and 12.5.



Supplementary Figure 9. Histone modifications and transcript levels of EC-associated genes. Median relative RPKM (background corrected) of H3K4me3 and H3K9ac (± 1 kb from TSS), H3K36me3 (gene bodies) and RNAseq (exons) for genes near ECs per heatmap cluster (Figure 3d).



Supplementary Figure 10. Maternal and zygotic control of embryonic transcripts. Maternally defined (MaD) peaks emerge at or before stage 11 independent of embryonic transcription. Zygotically defined (ZyD) peaks appear before stage 11 and are lost in α-amanitin treated embryos, or emerge at or after stage 12. Not determined (ND) peaks are not detected in stage 11 control embryos. **(a)** Maternal and zygotic control of H3K4me3 and H3K27me3 on promoters of embryonic transcripts (total number of transcripts: 983). **(b)** Box plots of p300 RPKM (background corrected) in GREAT regions of genes, or H3K27me3 RPKM (background corrected) in promoters of genes with at least one H3K27me3 peak in their promoter (± 2.5 kb from TSS). Box: 25th (bottom), 50th (internal band), 75th (top) percentiles. Whiskers: 1.5 * interquartile range of the lower and upper quartiles, respectively. Outliers are indicated with black dots. **(c)** BioCAP enrichment (RPKM BioCAP/ RPKM Input) as a measure for hypomethylated DNA domains on the promoters (± 1 kb from TSS) of early and late Wnt target genes.

REFERENCES

1. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553-560 (2007).
2. Strahl BD, Allis CD. The language of covalent histone modifications. Nature 403, 41-45 (2000).
3. Lee JS, Smith E, Shilatifard A. The language of histone crosstalk. Cell 142, 682-685 (2010).
4. Jenuwein T, Allis CD. Translating the histone code. Science 293, 1074-1080 (2001).
5. Gerstein MB, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330, 1775-1787 (2010).
6. Roy S, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-1797 (2010).
7. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development 136, 3033-3042 (2009).
8. Kane DA, Kimmel CB. The zebrafish midblastula transition. Development 119, 447-456 (1993).
9. O'Farrell PH, Stumpff J, Su TT. Embryonic cleavage cycles: how is a mouse like a fly? Current biology : CB 14, R35-45 (2004).
10. Paranjpe SS, Veenstra GJ. Establishing pluripotency in early development. Biochimica et biophysica acta, (2015).
11. Newport J, Kirschner M. A major developmental transition in early *Xenopus* embryos: I. characterization and timing of cellular changes at the midblastula stage. Cell 30, 675-686 (1982).
12. Newport J, Kirschner M. A major developmental transition in early *Xenopus* embryos: II. Control of the onset of transcription. Cell 30, 687-696 (1982).
13. van Heeringen SJ, et al. Principles of nucleation of H3K27 methylation during embryonic development. Genome research 24, 401-410 (2014).
14. Akkers RC, et al. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in *Xenopus* embryos. Developmental cell 17, 425-434 (2009).
15. Vastenhouw NL, et al. Chromatin signature of embryonic pluripotency is established during genome activation. Nature 464, 922-926 (2010).
16. Lindeman LC, et al. Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Developmental cell 21, 993-1004 (2011).
17. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182-187 (2010).
18. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854-858 (2009).
19. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108-112 (2009).
20. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics 39, 311-318 (2007).
21. Roh TY, Wei G, Farrell CM, Zhao K. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. Genome research 17, 74-81 (2007).
22. Bernstein BE, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120, 169-181 (2005).
23. Santos-Rosa H, et al. Active genes are tri-methylated at K4 of histone H3. Nature 419, 407-411 (2002).
24. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell 129, 823-837 (2007).
25. Schotta G, et al. A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes & development 18, 1251-1262 (2004).

26. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods 9, 215-216 (2012).
27. Chafin DR, Guo H, Price DH. Action of alpha-amanitin during pyrophosphorolysis and elongation by RNA polymerase II. The Journal of biological chemistry 270, 19114-19119 (1995).
28. Sible JC, Anderson JA, Lewellyn AL, Maller JL. Zygotic transcription is required to block a maternal program of apoptosis in *Xenopus* embryos. Developmental biology 189, 335-346 (1997).
29. Skirkanich J, Luxardi G, Yang J, Kodjabachian L, Klein PS. An essential role for transcription before the MBT in *Xenopus laevis*. Developmental biology 357, 478-491 (2011).
30. Clouaire T, et al. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes & development 26, 1714-1728 (2012).
31. Clouaire T, Webb S, Bird A. Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. Genome biology 15, 451 (2014).
32. Thomson JP, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464, 1082-1086 (2010).
33. van Kruijsbergen I, Hontelez S, Veenstra GJ. Recruiting polycomb to chromatin. The international journal of biochemistry & cell biology 67, 177-187 (2015).
34. Ng HH, Robert F, Young RA, Struhl K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Molecular cell 11, 709-719 (2003).
35. Paranjpe SS, Jacobi UG, van Heeringen SJ, Veenstra GJ. A genome-wide survey of maternal and embryonic transcripts during *Xenopus tropicalis* development. BMC genomics 14, 762 (2013).
36. Lauberth SM, et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152, 1021-1036 (2013).
37. Vermeulen M, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 131, 58-69 (2007).
38. Sims RJ, 3rd, et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Molecular cell 28, 665-676 (2007).
39. Gentsch GE, et al. In vivo T-box transcription factor profiling reveals joint regulation of embryonic neuromesodermal bipotency. Cell reports 4, 1185-1196 (2013).
40. Yasuoka Y, et al. Occupancy of tissue-specific cis-regulatory modules by Otx2 and TLE/Groucho for embryonic head specification. Nature communications 5, 4322 (2014).
41. Chiu WT, et al. Genome-wide view of TGFbeta/Foxh1 regulation of the early mesendoderm program. Development 141, 4537-4547 (2014).
42. Pott S, Lieb JD. What are super-enhancers? Nature genetics 47, 8-12 (2015).
43. Whyte WA, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307-319 (2013).
44. Hnisz D, et al. Super-enhancers in the control of cell identity and disease. Cell 155, 934-947 (2013).
45. Blythe SA, Cha SW, Tadjuidje E, Heasman J, Klein PS. beta-Catenin primes organizer gene expression by recruiting a histone H3 arginine 8 methyltransferase, Prmt2. Developmental cell 19, 220-231 (2010).
46. Bogdanovic O, et al. Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. Genome research 21, 1313-1327 (2011).
47. Potok ME, Nix DA, Parnell TJ, Cairns BR. Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell 153, 759-772 (2013).

48. Jiang L, et al. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell 153, 773-784 (2013).
49. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495-501 (2014).
50. Huether R, et al. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nature communications 5, 3630 (2014).
51. Wu G, et al. The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. Nature genetics 46, 444-450 (2014).
52. Simmer F, et al. Comparative genome-wide DNA methylation analysis of colorectal tumor and matched normal tissues. Epigenetics 7, 1355-1367 (2012).
53. Brinkman AB, et al. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome research 22, 1128-1138 (2012).
54. Jallow Z, Jacobi UG, Weeks DL, Dawid IB, Veenstra GJ. Specialized and redundant roles of TBP and a vertebrate-specific TBP paralog in embryonic gene regulation in *Xenopus*. Proceedings of the National Academy of Sciences of the United States of America 101, 13525-13530 (2004).
55. Akkers RC, Jacobi UG, Veenstra GJ. Chromatin immunoprecipitation analysis of *Xenopus* embryos. Methods in molecular biology 917, 279-292 (2012).
56. Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Gomez-Skarmeta JL. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. Methods 62, 207-215 (2013).
57. Lister R, et al. Global epigenomic reconfiguration during mammalian brain development. Science 341, 1237905 (2013).
58. Lister R, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68-73 (2011).
59. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology 9, R137 (2008).
60. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology 26, 1351-1359 (2008).
61. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome biology 13, R16 (2012).
62. Long HK, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. eLife 2, e00348 (2013).

CHAPTER THREE

fluff: exploratory analysis and visualization of high-throughput sequencing data

Georgios Georgiou
Simon J. van Heeringen

ABSTRACT

Summary

In this application note we describe fluff, a software package that allows for simple exploration, clustering and visualization of high-throughput sequencing data mapped to a reference genome. The package contains three command-line tools to generate publication-quality figures in an uncomplicated manner using sensible defaults. Genome-wide data can be aggregated, clustered and visualized in a heatmap, according to different clustering methods. This includes a predefined setting to identify dynamic clusters between different conditions or developmental stages. Alternatively, clustered data can be visualized in a bandplot. Finally, fluff includes a tool to generate genomic profiles. As command-line tools, the fluff programs can easily be integrated into standard analysis pipelines. The installation is straightforward and documentation is available at http://fluff.readthedocs.org.

Availability

fluff is implemented in Python and runs on Linux. The source code is freely available for download at https://github.com/simonvh/fluff.

INTRODUCTION

The advances in sequencing technology and the reduction of costs have led to a rapid increase of High- Throughput Sequencing (HTS) data. Applications include chromatin immunoprecipitation followed by high-throughput deep sequencing (ChIP-seq; Robertson et al. (2007)) to determine the genomic location of DNA-associated proteins, chromatin accessibility assays (Buenrostro et al., 2013; Hesselberth et al., 2009) and bisulfite sequencing to assay DNA methylation (Lister et al., 2009). The integration of these diverse data allow identification of the epigenomic state, for instance in different tissues (Martens and Stunnenberg, 2013; Roadmap Epigenomics Consortium et al., 2015) or during development (Hontelez et al., 2015). However, the scale and complexity of these datasets call for the use of computational methods that facilitate data exploration and visualization. Various options exist to explore and visualize HTS data mapped to a reference genome, for instance in aggregated form such as heatmaps and average profiles. These include general purpose modules for specific programming languages (Huber et al., 2015), dedicated HTS modules (Dale et al., 2014; Statham et al., 2010; Akalin et al., 2015), command-line tools (Shen et al., 2014; Giannopoulou and Elemento, 2011), web tools (Ram´ırez et al., 2014), stand-alone applications (Ram´ırez et al., 2014; Ye et al., 2011) and tools that depend on other software for visualization (Heinz et al., 2010). Here, we present fluff, a Python package for visual, reference-based HTS data exploration. It includes command-line applications to both cluster and visualize aggregated signals in genomic regions, as well as to create genome browser-like profiles. The scripts can be included in analysis pipelines and accept commonly used file formats. The fluff applications are pitched at the beginner to intermediate user. They have sensible defaults, yet allow for customizable creation of high-quality, publication-ready figures.

METHODS

General

Detailed documentation, including tutorials, is available at http://fluff.readthedocs.org. Fluff is imple- mented in Python and uses several previously published modules (Brewer (2016); Anders et al. (2015); Dale et al. (2011); Quinlan and Hall (2010); Li et al. (2009); de Hoon et al. (2004), see Supplemental Information). All fluff tools support indexed BAM, bigWig or (tabix-indexed) BED, WIG or bedGraph files as input. A large selection of major image formats are supported as output. The fluff tools were developed to explore ChIP-seq data, however, they will work with any type of data where (spliced) reads can be mapped to a genomic reference. For instance DNA methylation profiles from bisulfite-sequencing or RNA-seq data (Supplemental Figure 1) can also be visualized.

Normalization Normalization of sequencing data is critical for downstream analysis and various methods have been proposed (see for instance Angelini et al. (2015) and Bailey et al. (2013) for an overview of ChIP-seq normalization methods). For visualization, the most important factor is the sequencing read depth. Therefore, fluff has the option to normalize to the total number of mapped reads. Alternatively, averaged signal files such as bigWig tracks that are processed or normalized by a different method can be used as input.

Program descriptions

Heatmaps Visualization of HTS data as heatmaps, where rows represent different genomic regions, can highlight important aspects of the data, like differential enrichment or positional patterns for specific groups of features. In addition, it allows for comparison between multiple regions within the same or between different experiments. The fluff heatmap tool visualizes HTS data on basis of list of genomic coordinates. The data can optionally be clustered using either k-means or hierarchical clustering. For clustering, the read counts in the bins are normalized to the 75 percentile. The distance can be calculated using either the Euclidean distance or Pearson correlation similarity.

If the regions in the input file are not strand-specific, different clusters might represent the same strand-specific profile in two different orientations. Clusters that are mirrored relative to the center can optionally be merged. Here, the similarity is based on the chi-squared p-value of the mean profile per cluster. One important use case for clustering is the ability to identify dynamic patterns, for instance during different time points or conditions. For this purpose, clustering on the binned signal is not ideal. Therefore, fluffheatmap provides the option to cluster genomic regions based on a single value derived from the number of reads in the feature centers (+/- 1kb). In combination with the Pearson correlation metric, this allows for efficient retrieval of dynamic clusters. The difference is illustrated in Figure 2.

Bandplots In heatmaps, more subtle patterns can be difficult to detect, as the dynamic range of signal intensities is not well-reflected in the color scale. Therefore, as an alternative to a heatmap, fluff bandplot plots the average profiles in small multiples (Shoresh and Wong, 2012). Here, the spatial encoding of the signal allows for more accurate comparison of values (Gehlenborg et al., 2012). The median enrichment is visualized as a black line with the 50th and 90th percentile as a dark and light colour respectively.

Profiles Genome browsers are unrivalled for data exploration and visualization in a genomic context. However, it can be useful to create profiles of HTS data in genomic intervals using a consistent command- line tool, that can optionally be automated. The fluff profile tool

can plot summarized profiles from one or more profiles, together with (gene) annotation from a BED12-formatted file.

Analysis

In short, FASTQ files were downloaded from NCBI GEO (Edgar et al., 2002) and mapped to the human genome (hg19) using bwa (Li and Durbin, 2009). Duplicate reads were marked using bamUtil (http://genome.sph.umich.edu/wiki/BamUtil). All BAM files from replicate experiments were merged. Peaks were called using MACS2 (Zhang et al., 2008) with default settings. See Supplemental Information for specific details and accession numbers.

RESULTS

Demonstrating fluff: dynamic enhancers during macrophage differentiation

To illustrate the functionality of fluff we visualized previously published ChIP-seq data (Saeed et al., 2014). Here, the epigenomes of human monocytes and in vitro-differentiated naıve, tolerized, and trained macrophages were analyzed, with the aim to understand the epigenetic basis of innate immunity. Circulating monocytes (Mo) were differentiated into three macrophages states: to macrophages (Mf), to long-term tolerant cells (LPS-Mf) by exposition to lipopolysaccharide and to trained immune cells (BG-Mf) by priming with β-glucan. We used fluff heatmap to cluster and visualize the signal of histone 3 lysine 27 acetylation (H3K27ac), which is located at active enhancers and promoters (Fig. 1A). The input consisted of a BED file with 7,611 differentially regulated enhancers (Supplemental Table 1) and four BAM files, for each of the monocytes and three types of macrophages. Using k-means clustering (k = 5) with the Pearson correlation metric, the heatmap recapitulates the H3K27ac dynamics as described (Saeed et al., 2014).

While heatmaps are often used for visualization of signals over genomic features, either clustered or ordered by signal intensity, it can be difficult to distinguish relative levels of individual clusters. Figure 1B shows an alternative visualization of average enrichment profiles in small multiples. The same clusters as in Fig. 1A are plotted using fluffbandplot. Shown are the median (black line), along with the 50th (darker color) and 90th percentile (lighter color) of the data. This allows for more detailed comparisons.

Finally, we illustrate fluffprofile, which can visualize one or more genomic regions (Fig. 1C). This figure highlights the CNRIP1 gene from cluster 2, which shows a consistent increase of H3K27ac from Mo to Mf, LPS-Mf and BG-Mf. The signal profiles are directly generated from the BAM files.

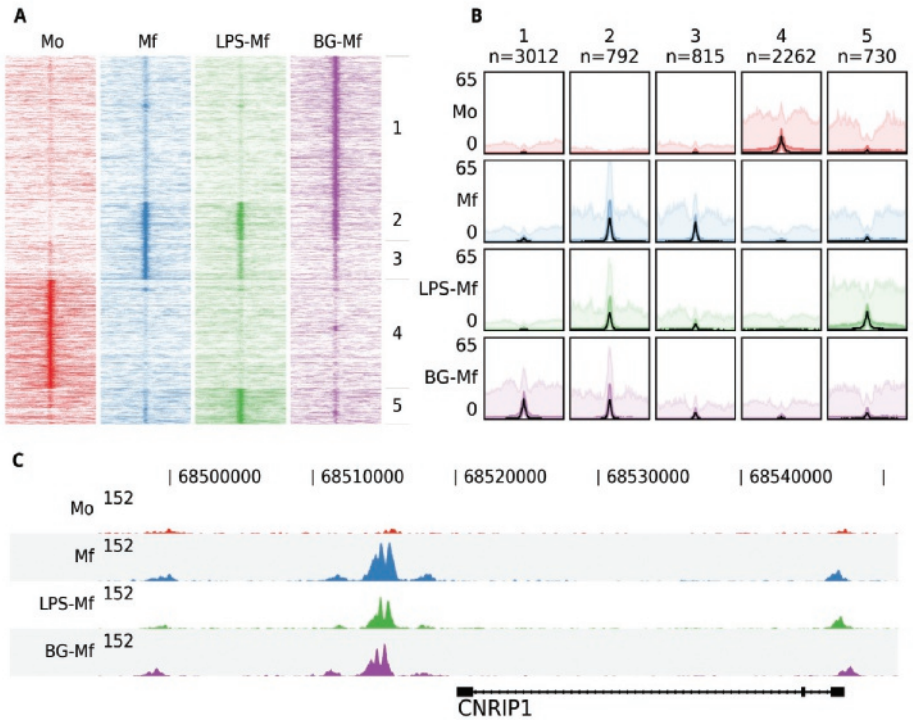


Figure 1. An example of the fluff output. All panels were generated by the fluff command-line tools and were not post-processed or edited. **(A)** Heatmap showing the results of k-means clustering (k=5, metric=Pearson) of dynamic H3K27ac regions in monocytes (Mo), na¨ıve macrophages (Mf), tolerized (LPS-Mf) and trained cells (BG-Mf) (Saeed et al., 2014). ChIP-seq read counts are visualized in 100-bp bins in 24-kb regions. **(B)** Bandplot showing the average profile (median: black, 50 percent: dark color, 90 percent: light color) of the clusters as identified in Fig. 1A. **(C)** The H3K27ac ChIP-seq profiles at the CNRIP1 gene locus, which shows a gain of H3K27ac in Mf, LPS-Mf and BG-Mf relative to Mo.

Identification and visualization of dynamic patterns

Most applications that cluster HTS data for heatmap visualization use a binning approach, followed by clustering using the Euclidean distance. The implicit effect is that the bins are clustered on basis of the spatial patterns relative to the region of interest. Often, this is the desired result, for instance when clustering the ChIP-seq enrichment patterns of different histone modifications at the transcription start sites of genes. However, for other analyses this clustering approach does not suffice. An example could be the ChIP-seq profiles of specific histone modifications correlated to the activity of a regulatory element, such as H3K4me3 at promoters or H3K27ac at enhancers. In this case, a relevant objective is to identify the clusters associated with differential activation dynamics. As illustration, we visualized the H3K27ac enrichment profile at DNaseI hypersensitive sites in human embryonic stem (ES) cells

differentiated into different lineages (Xie et al., 2013). Here, H1 ES cells were differentiated into mesendoderm, neural progenitor cells, trophoblast-like cells, and mesenchymal stem cells. We first clustered the H3K27ac profiles at regulatory elements on chromosome 1 using the standard approach, based on comparing all the bins using the Euclidean distance metric (Fig. 2A).

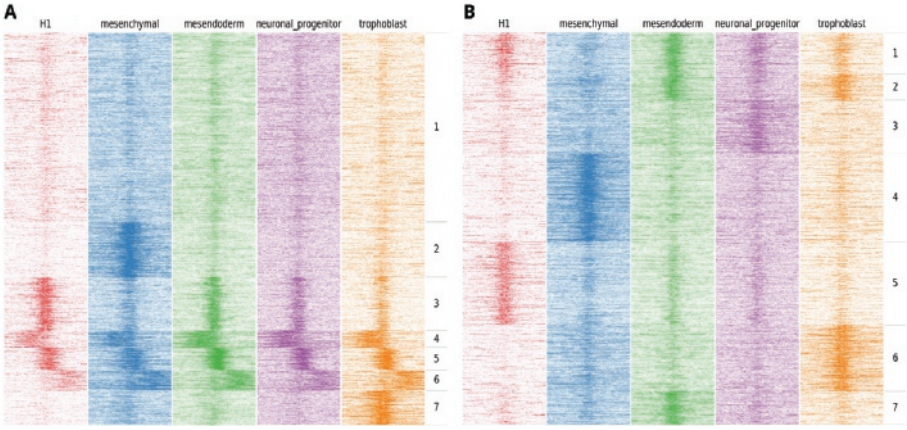


Figure 2. Example of the output of fluffheatmap using standard clustering compared to using the dynamics option. Shown are the H3K27ac ChIP-seq read counts in 100bp bins in 20kb around the DNaseI peak summit in human H1 ES cell-derived cells. **(A)** Heatmap showing the results of k-means clustering of all bins (k=7, metric=Euclidean) **(B)** Heatmap showing the results of k-means clustering in 2kb regions centered at the peak summit (k=7, metric=Pearson).

Here, we identify two clusters with high enrichment (cluster 3 and cluster 5), a cluster with relatively low, narrow enrichment (cluster 1), and two clusters with broad enhancer domains (cluster 4 and 6). However, only two strong dynamic clusters are identified, cluster 2, which shows enhancers specifically activated in mesenchymal stem cells and cluster 7 which shows enhancers specifically activated in trophoblast-like stem cells. Figure 2B shows an alternative clustering approach implemented in fluff heatmap. Here the regions were clustered on basis of the Pearson correlation of read counts in the center of the region (extended to 2kb). This shows a completely different picture and we now can identify enhancers specific to H1 ES cells (cluster 5), mesenchymal (cluster 4), mesendoderm (cluster 7), neuronal progenitor (cluster 3) and trophoblast cells (cluster 6). These lineage-specific enhancer dynamics were not visible in the clustering in Figure 2A.

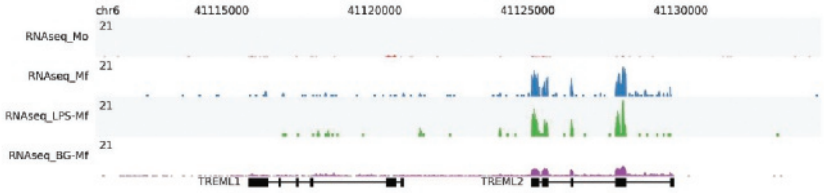
CONCLUSION

The analysis of multi-dimensional genomic data requires methods for data exploration and visualization. We provide fluff, a Python package that contains several command-line tools to generate figures for use in high-throughput sequencing analysis workflows. We aim to fill the gap between powerful, flexible libraries that require programming skills on the one hand, and intuitive, graphical programs with limited customization possibilities on the other hand. These tools were developed based on a need for straight- forward analysis and visualization of ChIP-seq data and have been successfully applied in a variety of projects (Menafra et al., 2014; van den Boom et al., 2016; Kouwenhoven et al., 2015). In conclusion, fluff helps to interpret genome-wide experiments by efficient visualization of sequencing data.

ACKNOWLEDGEMENTS

This study makes use of data generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Additionally, this study used data provided by the NIH Roadmap Epigenomics Consortium (http://nihroadmap.nih.gov/epigenomics/) .

SUPPLEMENTARY FIGURES



Supplemental Figure 1. Visualization of RNAseq data using fluffprofile. Shown are the RNA-seq profiles at the TREML1 and TREML2 gene loci of Monocytes (red), Macrophages (blue), Macrophages preincubated with LPS (green) and Macrophages preincubated with βglucan (purple). Read depth (per million reads) is normalized to the total number of mapped reads per sample.

SUPPLEMENTARY METHODS

Implementation

The fluff module and commandline tools are implemented in Python and make use of the following packages:

- colorbrewer (Brewer, 2016)
- HTSeq (Anders et al., 2015)
- pybedtools (Dale et al., 2011; Quinlan and Hall, 2010)
- pysam (Li et al. (2009); pysam htslib interface for python)
- pycluster (de Hoon et al., 2004)

In addition, fluff uses the numpy, scipy and matplotlib Python libraries.

The package can be installed using the Python package manager pip or the conda package manager from the Anaconda open source analytics platform (https://continuum.io). The source code is freely available at http://github.com/simonvh/fluff under a MIT license.

Data description

The H3K27ac ChIPseq and RNAseq data that were used for the monocytemacrophage analysis (Fig. 1; Fig. S1) were downloaded from NCBI GEO (Edgar et al., 2002), accession GSE58310, and are described in Saeed et al. (2014). The specific samples are listed in Table 1. The ChIPseq FASTQ files were mapped to the human genome (hg19) using bwa version 0.7.10 (Li and Durbin, 2009). The RNAseq FASTQ files were mapped to the human genome (hg19) using gsnap version 20120720 (Wu and Nacu 2010). Duplicate reads were marked using bamUtil 1.0.2 (http://genome.sph.umich.edu/wiki/BamUtil). The BAM files were filtered to remove all reads

with mapping quality less than 15. The regions that were used as input for Figs. 1A and B are supplied in Supplementary Table 2.

The H3K27ac ChIPseq and DNaseI data in human H1 cells (Fig. 2) were downloaded from NCBI GEO (Edgar et al., 2002), series accessions GSE18927 and GSE16256, and are described in Xie et al. (2013).. The FASTQ files were mapped to the human genome (hg19) using bwa version 0.7.12r1039 (Li and Durbin, 2009). Duplicate reads were marked using bamUtil 1.0.2 (http://genome.sph.umich.edu/wiki/BamUtil). All BAM files from replicate experiments were merged. Peaks were called on the DNaseI BAM files using MACS2 2.1.0.20140616 (Zhang et al. (2008); https://github.com/taoliu/MACS/) with default settings. All DNaseI peaks of different experiments were merged and centered on the highest summit as determined by MACS2. The peaks and reads corresponding to chromosome 1 were filtered. This data set is available from figshare (van Heeringen, 2016; DOI:10.6084/m9.figshare.3113728.v1).

Command lines

To create the panels for Figure 1, Figure 2 and Supplementary Figure 1, fluff was run with the following settings:

Fig. 1A:

```
fluff heatmap -f dynamic_regions.bed -d Mo.bam Mf.bam LPS-Mf.bam BG-Mf.bam -C kmeans
-k5 -M pearson -g -e 12000 -T 20 -o Saeed_dynamicRegions_Pearson_K5_e12000_g_T20
```

Fig. 1B:

```
fluff bandplot -f Saeed_dynamicRegions_Pearson_K5_e12000_g_clusters.bed -counts
Saeed_dynamicRegions_Pearson_K5_e12000_g_readCounts.txt -s 1:4 -P 98.5 –T 20  -o
Saeed_dynamicRegions_Pearson_K5_e12000_g_T20_bandplot_T20
```

Fig. 1C:

```
fluff profile -I chr2:68495000-68551000 -d Mo.bam Mf.bam LPS-Mf.bam BG-Mf.bam -s 1:4 -T
10 -a hg19_geneAnnotation.bed -o CNRIP1_profile_chr2_68495000_68551000_T10
```

Fig. 2A:

```
fluff heatmap -f example_peaks.bed -d H1.bammesenchymal.bam mesendoderm.bam
neuronal_progenitortrophoblast.bam -C k -k 7 -o H3K27ac_kmeans7-P5
```

Fig. 2B:

```
fluff heatmap-fexample_peaks.bed -d H1.bammesenchymal.bam mesendoderm.bam
neuronal_progenitortrophoblast.bam -C k -k 7 -g -M p -o H3K27ac_kmeans7_dynamics
```

Fig. S1:

```
fluff profile -l chr6:41112015-41135714 -d RNAseq_Mo.bam RNAseq_Mf.bam RNAseq_LPS-Mf.bam RNAseq_BG-Mf.bam -a hg19_geneAnnotation.bed -f 0 -s 1:4 -n -o RNAseq_TREML_chr6_41112015_41135714_f0_normalized
```

REFERENCES

Akalin, A., Franke, V., Vlahovi˘cek, K., Mason, C. E., and Sch¨ubeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics, 31(7):1127–1129.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a python framework to work with high-throughput sequencing data. Bioinformatics, 31(2):166–169. Angelini, C., Heller, R., Volkinshtein, R., and Yekutieli, D. (2015). Is this the right normalization? a diagnostic tool for ChIP-seq normalization. BMC Bioinformatics, 16:150.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput. Biol., 9(11):e1003326.

Brewer, C. (2016). ColorBrewer: Color advice for maps. http://www.colorbrewer2.org, accessed: 2016-3-15. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods, 10(12):1213–1218.

Dale, R. K., Matzat, L. H., and Lei, E. P. (2014). metaseq: a python package for integrative genome-wide analysis reveals relationships between chromatin insulators and associated nuclear mRNA. Nucleic Acids Res., 42(14):9158–9170.

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics, 27(24):3423–3424.

de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics, 20(9):1453–1454.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res., 30(1):207–210. Gehlenborg, N., Nils, G., and Bang, W. (2012). Points of view: Heat maps. Nat. Methods, 9(3):213–213. Giannopoulou, E. G. and Elemento, O. (2011). An integrated ChIP-seq analysis platform with customiz- able workflows. BMC Bioinformatics, 12:277.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell, 38(4):576–589. Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods, 6(4):283–289. Hontelez, S., van Kruijsbergen, I., Georgiou, G., van Heeringen, S. J., Bogdanovic, O., Lister, R., and Veenstra, G. J. C. (2015). Embryonic transcription is controlled by maternally defined chromatin state. Nat. Commun., 6:10148.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole´s, A. K., Pag`es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. Nat. Methods, 12(2):115–121.

Kouwenhoven, E. N., Oti, M., Niehues, H., van Heeringen, S. J., Schalkwijk, J., Stunnenberg, H. G., van Bokhoven, H., and Zhou, H. (2015). Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. EMBO Rep., 16(7):863–878.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078–2079.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271):315–322. Martens, J. H. A. and Stunnenberg, H. G. (2013). BLUEPRINT: mapping human blood cell epigenomes. Haematologica, 98(10):1487–1489. Menafra, R., Brinkman, A. B., Matarese, F., Franci, G., Bartels, S. J. J., Nguyen, L., Shimbo, T., Wade, P. A., Hubner, N. C., and Stunnenberg, H. G. (2014). Genome-wide binding of MBD2 reveals strong preference for highly methylated loci. PLoS One, 9(6):e99603. Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841–842.

Ram´ırez, F., D¨undar, F., Diehl, S., Gr¨uning, B. A., and Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res., 42(Web Server issue):W187–91. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi- Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539):317–330. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods, 4(8):651–657.

Saeed, S., Quintin, J., Kerstens, H. H. D., Rao, N. A., Aghajanirefah, A., Matarese, F., Cheng, S.-C., Ratter, J., Berentsen, K., van der Ent, M. A., Sharifi, N., Janssen-Megens, E. M., Ter Huurne, M., Mandoli, A., van Schaik, T., Ng, A., Burden, F., Downes, K., Frontini, M., Kumar, V., Giamarellos-Bourboulis, E. J., Ouwehand, W. H., van der Meer, J. W. M., Joosten, L. A. B., Wijmenga, C., Martens, J. H. A., Xavier, R. J., Logie, C., Netea, M. G., and Stunnenberg, H. G. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. Science, 345(6204):1251086.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs. plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics, 15:284. Shoresh, N. and Wong, B. (2012). Points of view: Data exploration. Nat. Methods, 9(1):5. Statham, A. L., Strbenac, D., Coolen, M. W., Stirzaker, C., Clark, S. J., and Robinson, M. D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. Bioinformatics, 26(13):1662–1663.

van den Boom, V., Maat, H., Geugien, M., Rodr´iguez L´opez, A., Sotoca, A. M., Jaques, J., Brouwers-Vos, A. Z., Fusetti, F., Groen, R. W. J., Yuan, H., Martens, A. C. M., Stunnenberg, H. G., Vellenga, E., Martens, J. H. A., and Schuringa, J. J. (2016). Non-canonical PRC1.1 targets active genes independent of H3K27me3 and is essential for leukemogenesis. Cell Rep., 14(2):332–346.

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C.-A., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R., and Ren, B. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell, 153(5):1134–1148.

Ye, T., Krebs, A. R., Choukrallah, M.-A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res., 39(6):e35.

Zhang, Y., Yong, Z., Tao, L., Meyer, C. A., J´erˆome, E., Johnson, D. S., Bernstein, B. E., Chad, N., Myers, R. M., Myles, B., Wei, L., and Shirley Liu, X. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9(9):R137.

CHAPTER FOUR

Regulatory remodeling in the allo-tetraploid frog *Xenopus laevis*

Dei M. Elurbe [*]
Sarita S. Paranjpe [*]
Georgios Georgiou [*]
Ila van Kruijsbergen
Ozren Bogdanovic
Romain Gibeaux
Rebecca Heald
Ryan Lister
Martijn A. Huynen
Simon J. van Heeringen
Gert Jan C. Veenstra

ABSTRACT

Background

Genome duplication has played a pivotal role in the evolution of many eukaryotic lineages, including the vertebrates. A relatively recent vertebrate genome duplication is that in *Xenopus laevis*, which resulted from the hybridization of two closely related species about 17 million years ago. However, little is known about the consequences of this duplication at the level of the genome, the epigenome, and gene expression.

Results

The *X. laevis* genome consists of two subgenomes, referred to as L (long chromosomes) and S (short chromosomes), that originated from distinct diploid progenitors. Of the parental subgenomes, S chromosomes have degraded faster than L chromosomes from the point of genome duplication until the present day. Deletions appear to have the largest effect on pseudogene formation and loss of regulatory regions. Deleted regions are enriched for long DNA repeats and the flanking regions have high alignment scores, suggesting that non-allelic homologous recombination has played a significant role in the loss of DNA. To assess innovations in the *X. laevis* subgenomes we examined p300-bound enhancer peaks that are unique to one subgenome and absent from *X. tropicalis*. A large majority of new enhancers comprised of transposable elements. Finally, to dissect early and late events following interspecific hybridization, we examined the epigenome and the enhancer landscape in *X. tropicalis* × *X. laevis* hybrid embryos. Strikingly, young *X. tropicalis* DNA transposons are derepressed and recruit p300 in hybrid embryos.

Conclusions

The results show that erosion of *X. laevis* genes and functional regulatory elements is associated with repeats and non-allelic homologous recombination and furthermore that young repeats have also contributed to the p300-bound regulatory landscape following hybridization and whole-genome duplication.

BACKGROUND

Genome duplication is a major force in genome evolution that not only doubles the genetic material but also facilitates morphological innovations. In plants, whole-genome duplications (WGD) appear to occur more often than in animals [1] and some phenotypic innovations, like the origin of flowers, have been attributed to this phenomenon [2]. In animals, two rounds of WGD at the root of the vertebrate tree (~ 500 million years ago [Mya]) gave rise to the four HOX clusters and have led to the expansion of the neural synapse proteome [3]. It is likely that this facilitated an increase in the morphological complexity [4] and allowed an increase in the complexity in the vertebrate behavioral repertoire [5]. More recent genome duplications have been documented in fish, at the root of the teleost fish 320 Mya and in the common ancestor of salmonids 80 Mya [6]. Amphibians in general appear to have undergone many polyploidizations, with natural polyploids in 15 Anuran and in four Urodelan families. In *Xenopus* (African clawed frogs), duplications have occurred on multiple occasions, giving rise to tetraploid, octoploid, and dodecaploid species [7]. One such duplication occurred in the ancestor of the amphibian *Xenopus laevis* 17 Mya [8]. The allo-tetraploid genome of *X. laevis* consists of two subgenomes, referred to as L (long chromosomes) and S (short chromosomes), that originated from distinct diploid progenitors [8]. Most of the additional genes that result from WGD events tend to be lost in evolution. In the case of allopolyploidy, this loss is biased to one of the parental subgenomes [9], a phenomenon referred to as biased fractionation. One explanation for biased fractionation is the variation in the level of gene expression between the homeologous chromosomes [10], with the lowest expressed gene having the highest probability of being lost because it would contribute less to fitness.

The effects of polyploidization on the epigenome have mainly been studied in plants, where correlations between the gene expression and epigenetic modifications have been observed between homeologous genes [11], but are not well characterized in animals. The epigenetic modifications found in chromatin (DNA methylation and post-translational modifications of histones) are involved in gene regulation during development and differentiation [12], [13]. A high density of methylated CpG dinucleotides is repressive towards transcription; conversely, the DNA of a large fraction of promoters is unmethylated. In addition, histone H3 in promoter-associated nucleosomes is tri-methylated on lysine 4 (H3K4me3) when the promoter is active. Active enhancers on the other hand are decorated with mono-methylated H3K4 (H3K4me1) and they also recruit the p300 (*Ep300*) co-activator which can acetylate histones. When genes are expressed, they not only recruit RNA polymerase II (RNAPII), responsible for the production of the messenger RNA, but the gene body will be decorated with H3K36me3, which is left in the wake of elongating RNAPII. Therefore, deep sequencing approaches to determine these biochemical properties in a given tissue or developmental stage can be used to interrogate

the activity of genomic elements. This is highly relevant in the context of genomic evolution, as changes in gene expression caused by mutations in cis-regulatory elements are a major source of morphological change during evolution [14].

Here we ask how genome evolution and the epigenetic control of gene expression are related to interspecific hybridization and WGD. We compare functional regulatory elements in the L and S subgenomes of *X. laevis* embryos by chromatin immunoprecipitation (ChIP)-sequencing (ChIP- seq) of histone modifications, RNA-sequencing (RNA-seq), and whole genome bisulfite sequencing (WGBS) and use *Xenopus tropicalis*, a closely related diploid species, as a reference. We quantify the loss and the gain of genetic material and analyze how it has affected genes and gene-regulatory regions. Although genome evolution after the hybridization appears dominated by sequence loss, we also find evidence for the gain of functional elements. We specifically identify new subgenome-specific regulatory elements that recruit p300 and show that these are enriched for transposable elements (TEs). Finally, to assess the early gene-regulatory effects of hybridization we analyze experimental interspecific *X. tropicalis* × *X. laevis* hybrids and we observe hybrid-specific p300 recruitment to DNA transposons, further highlighting the role of such elements in the evolution of gene regulation.

RESULTS

The *X. laevis* L and S subgenomes show a bias in chromatin state and gene expression

To study the evolution of gene regulation in the context of WGD we generated transcriptomic and epigenomic profiles in *X. laevis* early gastrula embryos (Nieuwkoop-Faber stage 10.5; Additional file 1). We performed RNA-seq and obtained epigenomic profiles using ChIP followed by deep sequencing (ChIP-seq). We generated ChIP-seq profiles for H3K4me3, associated with promoters of active genes, H3K36me3, associated with actively transcribed genes, the Polr2a subunit of RNA Polymerase II (RNAPII), and the transcription coactivator p300. In addition, we performed WGBS to obtain DNA methylation profiles [15]. The sequencing results and details are summarized in additional file 1.

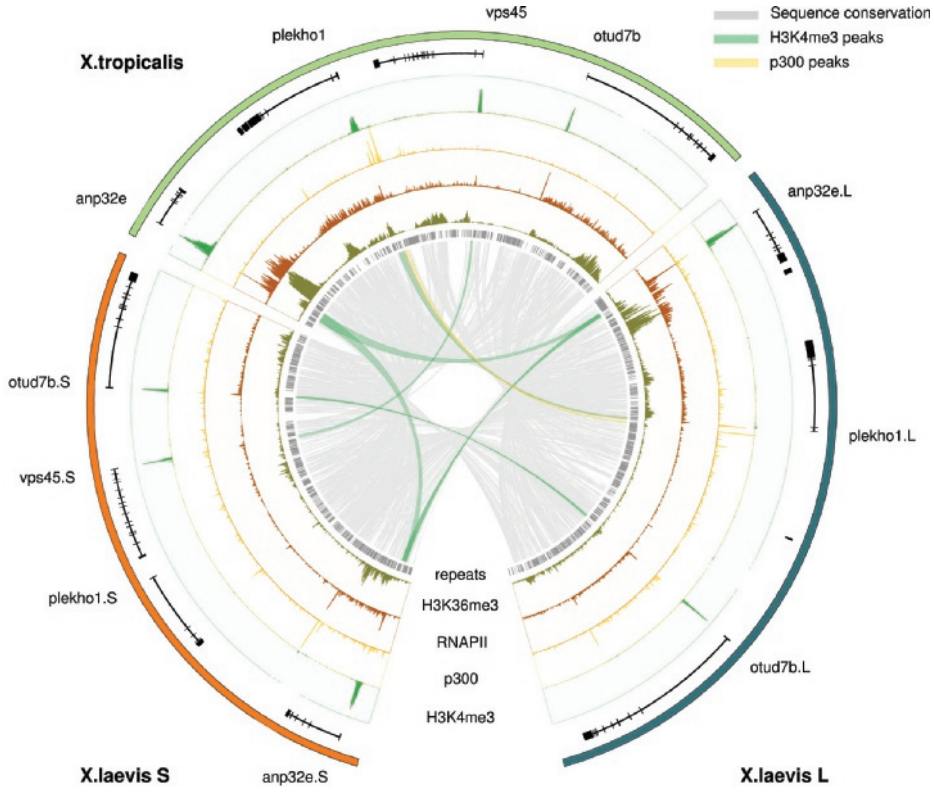


Figure 1. Alignment of a region on chromosome 8 in *X. tropicalis* and the *X. laevis* L and S subgenomes annotated with experimental ChIP-seq data (gastrula-stage embryos; NF stage 10.5). Shown are the gene annotation (black), repeats (gray), ChIP-seq profiles for H3K4me3 (green), p300 (yellow), RNA Polymerase II (RNAPII; brown), and H3K36me3 (dark green). The sequence conservation is indicated by gray lines. Conserved H3K4me3 and p300 peaks are denoted by green and yellow lines, respectively. The *anp32e* gene is expressed in *X. tropicalis* and both the L and S subgenome of *X. laevis*. The *plekho1* gene, on the other hand, has lost promoter and enhancer activity on the *X. laevis* S locus, and shows no experimental evidence of being expressed.

We created whole genome alignments (see Methods) to establish a framework for analysis of the epigenetic modifications in the two *X. laevis* subgenomes and in the *X. tropicalis* genome. Of the *X. laevis* L and S non- repetitive sequence, 61% and 59%, respectively, can be aligned with the orthologous *X. tropicalis* sequence. This allows for comparisons of the activity of genes and regulatory elements between homeologous regions. Figure 1 shows a region on *X. tropicalis* chromosome 8 containing four genes, together with the corresponding aligning sequences on chr8L and chr8S in *X. laevis*. The epigenomic profiles (H3K4me3, p300, RNAPII and H3K36me3) of both *X. laevis* and *X. tropicalis* [16] are shown and the sequence conservation obtained from the whole gene alignment is illustrated by gray lines in the center

of the plot. Regions that are conserved at both the sequence level and at the functional level (as measured by ChIP-seq) are highlighted. The *anp32e* gene is an example of a conserved gene that is expressed from all three genomes, as evidenced by H3K4me3 at the promoter and H3K36me3 and elongating RNAPII in the gene body. In contrast, expression of the *plekho1* gene has been lost from S. The gene is still present, but it is not active. There is no evidence of expression and both the H3K4me3 and the p300 signal are lost. Finally, the *vps45* gene is an example of a gene that is completely lost from L.

Next, we quantified gene expression patterns in the *X. laevis* subgenomes. Of the 17,303 genes expressed at stage 10.5, 9,230 can be assigned to the L subgenome and 6,685 to S. Of those expressed genes, 4,972 are singletons located on L and 2,646 on S. As reported previously [8], when both genes of a homeologous pair have detectable expression (3,545 genes), the expression level is correlated (Pearson R = 0.60, p <1e-300; Fig. 2a) and a minor but significant expression bias is detected (median expression difference of L compared to S = 5.7%; p < 1e-4; Wilcoxon signed-rank test). However, for many homeologs the expression bias is quite high, such that for one copy hardly any expression can be detected. Such non-expressed homeologs are located on both L and S, but occur more frequently on S (L:494, S:713; P=6.0e-11, Fisher's exact).

We examined whether the expression differences between the L and S homeologs could be explained by differential transcription regulation. We used the epigenomic profiles to assay the promoter state (H3K4me3, DNA methylation), enhancer activity (p300), and active expression (RNAPII, H3K36me3). The L subgenome has 38% more annotated genes than the S subgenome [8]. We observe the same trend for the regulatory elements. The number of H3K4me3 peaks, DNA-methylation free regions (see Methods) and p300 peaks is higher on L (28%, 23% and 35%, respectively; Additional file 2). The overall effect is that there is no significant difference between the numbers of regulatory elements per gene for the two subgenomes.

To analyze the conservation of regulatory elements, we compared the H3K4me3 and p300 data to similar ChIP-seq profiles from *X. tropicalis* obtained at the equivalent developmental stage [16]. In general promoters are much more conserved than enhancers (Fig. 2b). From all H3K4me3 peaks in *X. tropicalis*, ~40% are conserved in *X. laevis*, while for the p300 peaks the conservation is only ~13% (p < 1e-4; Chi-squared test). This is congruent with the finding in mammals that enhancers evolve much more rapidly than promoters [17]. Whereas the number of conserved regulatory elements is lower in S than in L, the elements that can be aligned differ relatively little at the sequence level and show over ~60% sequence identity (Fig. 2c).

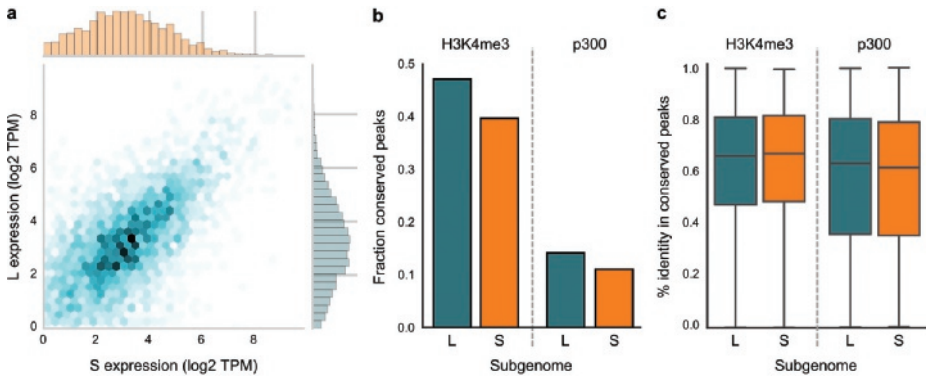


Figure 2. (a) Scatterplot of the expression level (log2 TPM) of L and S homeologs that are both expressed. The expression level of homeolog genes is generally similar (Pearson R = 0.60, p < 1e-300). **(b)** Fraction of epigenetic signals ("peaks") conserved in *X. laevis* compared to *X. tropicalis*. Promoters appear more conserved than enhancers; S has lost more epigenetic elements than L. **(c)** Active functional elements are equally conserved between L and S as compared to *X. tropicalis*. The background level of sequence conservation in fourfold degenerate sites from coding sequences with respect to *X. tropicalis* is 78.4% in L and 77.7% in S.

These analyses show that the L and S subgenomes have evolved differently with respect to gene content [8] and regulatory elements. Many more genes from S are lower expressed than their homeologs in L than vice versa. The number of functional regulatory elements, as identified by H3K4me3 and p300 ChIP-seq, is proportional to a more profound loss of homeologous genes from the S subgenome. Next, we set out to determine the origin of this differential loss.

Large deletions are prominent in the S subgenome

The chromosomes of the *X. laevis* S subgenome are substantially shorter than the L chromosomes. The average size difference is 17.3% based on the assembled sequence [8] and 13.2% based on the karyotype [18]. To investigate the cause of these differences, we analyzed the pattern of deletions on both subgenomes. We called deleted regions based on the absence of conservation between the *X. laevis* subgenomes if they were at least partly conserved between one *X. laevis* subgenome and *X. tropicalis*. In addition, to be able to measure the size of the deletions, we required that the putative deleted regions were flanked on both sides by conserved sequences on both *X. laevis* subgenomes (Additional file 3: Figure S1). This resulted in a set of 19,109 deletions, of which 13,066 (68%) were deleted from S (LΔS) and 6,043 (32%) were deleted from L (SΔL). There is a clear deletion bias towards S, which increases with the size of the deletion (Fig. 3a). These deletions affect genes and their regulatory sequences, as for example in the glrx2 locus where the promoter and most of the exons have been lost from the S subgenome (Fig. 3b). We asked to what extent functional sequences in the

L and S subgenomes are preserved (i.e. subject to fewer deletions) relative to the subgenome-specific deletion rates. To do that we randomly redistributed the deletions per chromosome and compared the effect on various annotated and experimentally derived features. As we cannot assess these features prior to their deletion we used the annotation and experimental data of the homeologous feature from the other subgenome as a proxy for the state in the genome from which that feature was deleted. The fold difference between the observed number of deleted basepairs and the expected number (mean of 1,000 randomizations) is visualized in Figure 3c. As expected, the frequency of deletions in intergenic regions and introns is similar relative to a uniform chromosomal distribution of deletions. The observed loss of exons on L is significantly lower than this randomized distribution ($p = 1.8e-20$; Fig. 3c). The fraction of exonic sequence that has disappeared is ~4-fold less than intronic or intergenic sequence (Additional file 3: Figure S2). This is likely the result of negative selection against loss. By contrast, for subgenome S the fraction of exonic sequence that has been deleted is similar to the rest of S (Fig. 3c) and exonic sequences in S appear not to be under selection against deletion. To obtain more direct evidence of functional sequences, we examined the loss of genomic elements that are decorated with RNAPII and the active transcription histone mark H3K36me3 (IntronicTx, ExonicTx, see Methods), with the enhancer coactivator p300, or with the active promoter mark H3K4me3. There appears to be strong selection on both S and L against deletion of actively transcribed exons (Fig. 3c, middle panel; $p = 2.4e-4$ and $p = 2.3e-7$, respectively) but not of transcribed introns. Furthermore, active enhancers and promoters in S and in L have significantly fewer deletions compared to the uniform chromosomal distribution (Fig. 3c; $p = 8.4e-7$, $p = 8.4e-8$, $p = 1.4e-5$ and $p = 2.9e-12$, respectively) and therewith appear to be under selection against loss. There is a large difference in the number of deletions between L and S (Fig. 3a), however, this in itself is not necessarily the result of selection as it mostly affects non-functional sequences (Fig. S2a). We asked if, on top of this difference in absolute number, there is evidence for more selection against deletions in L than in S. We therefore compared the reduction in the loss of transcribed exons, promoters and p300 elements relative to background loss between L and S. For all three the reduction in L appears to be larger than in S (Fig. 3c). For p300-bound enhancers and for H3K4me3-decorated promoters this difference in the reduction between L and S is significant ($p = 0.003$ and $p = 0.001$, respectively). This suggests that, aside from a higher deletion rate in S, there is also less selection against deletion of functional genetic elements in S than in L.

One of the possible sources of the loss of genomic DNA in the L and S subgenomes is non-allelic homologous recombination (NAHR), which is known to occur between long repetitive elements on the same chromosome [19]. To test whether this phenomenon could be responsible for the genomic losses detected, we examined the length distribution of repetitive elements in retained regions, i.e. the homeologous regions of the sequences that were lost in

one of the subgenomes (Fig. 3d). Indeed, we observe that repetitive elements are on average 3.7 times longer ($p < 1e-52$; Mann-Whitney U test) compared to random genomic sequences (Fig. 3d). Furthermore, the flanks of the retained regions (L for LΔS and S for SΔL, respectively) tend to be more similar to each other than random genomic sequences ($p < 1e-83$; Mann-Whitney U test; Fig. 3e). Nevertheless, the current density of repetitive elements is similar in the L and S subgenomes (Additional file 3: Figure S3), indicating that repeat density alone does not cause biased sequence loss on S chromosomes. These observations suggest that NAHR of ancient repeats has played a significant role in the deletions of regions from both subgenomes; the overall sequence loss is much more prevalent on the S chromosomes (Fig. 3a). To estimate when in the evolution these deletions and other types of mutations occurred we dated the origin of the pseudogenes that they caused.

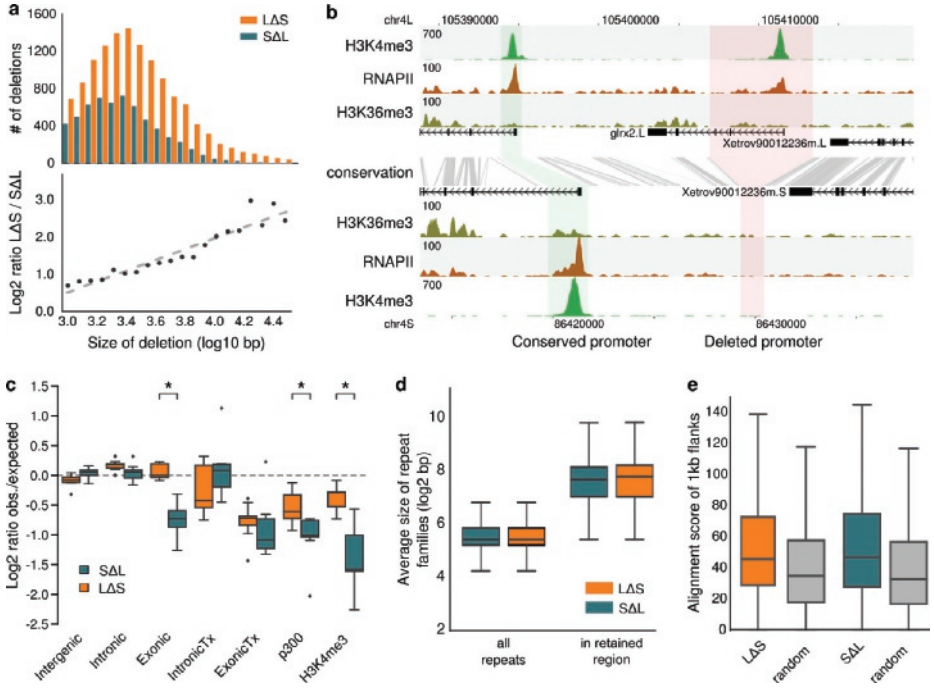


Figure 3. The S subgenome has more and larger deletions than L. (a) Size frequency distribution of deletions (top panel) and size ratio of LΔS deletions relative to SΔL deletions as a function of deletion size (bottom). **(b)** An example of a gene (*grlx2*) that has lost the promoter on the S genome due to a deletion. Shown are the gene annotation (black), ChIP-seq profiles for H3K4me3 (green), RNAPII (brown), and H3K36me3 (dark green). The sequence conservation is indicated by gray lines. **(c)** The log2 fold difference between the observed number of deleted basepairs and the expected number (mean of 1,000 randomizations). The fold difference is calculated per chromosome and summarized in a boxplot. Intergenic: 1kb distance from a gene. Intronic: introns.

Exonic: UTRs + CDS. IntronicTx: introns from genes actively transcribed. ExonicTx: Exons from genes actively transcribed. p300: genomic fragments having a p300 peak. H3K4me3: genomic fragments having a H3K4me3 peak. The asterisks mark significant differences between the L and S chromosomes ($p < 0.001$, Mann-Whitney U test). **(d)** Retained regions associated with deletions are enriched for relatively long repeats ($p < 1e\text{-}52$ for both LΔS and SΔL; Mann-Whitney U test) (e) 1kb flanks of the retained regions are more similar to each other than random genomic regions of the same size ($p < 1e\text{-}114$ and 1e-83 for LΔS and SΔL respectively; Mann-Whitney U test).

High levels of pseudogenization started after hybridization and continue to the present

To date the pseudogenes, we aligned them with the protein coding regions in L, S and the outgroup *X. tropicalis* (Methods: Search and alignment of orthologs and evolution rates). The coding regions in S are generally less conserved than in L, especially regarding synonymous substitutions (Ks, Fig. 4a, $p < 2.2e\text{-}16$; Wilcoxon signed-rank test). However, the ratio between nonsynonymous and synonymous substitutions (Ka/Ks) is only slightly higher in S compared to L (Fig. 4b, $p < 2.2e\text{-}16$; Wilcoxon signed-rank test). The difference in Ks between the L and S subgenomes shows that S has been subject to moderately higher mutation rates than L. In order to examine whether the relatively high level of mutations in the S genome persists to this day, we examined the level of SNPs separating the published inbred genome [8] and the progeny of two outbred individuals (Methods: SNP calling). We observe that the level of SNPs in the S genome is 3% higher than in the L genome in intergenic ($p = 5e\text{-}136$; Chi-squared test) and intronic regions ($p = 8e\text{-}101$; Chi-squared test). A similar difference is observed in 4-fold degenerate (4D) positions of coding DNA (also assumed to be under relaxed constraint) but this is not statistically significant (Additional file 4). The 4D positions exhibit a SNP density higher than in non-coding DNA; this correlates with an overrepresentation of CpGs in coding DNA (Additional file 3: Figure S4) and has been observed before in human genomes [20].

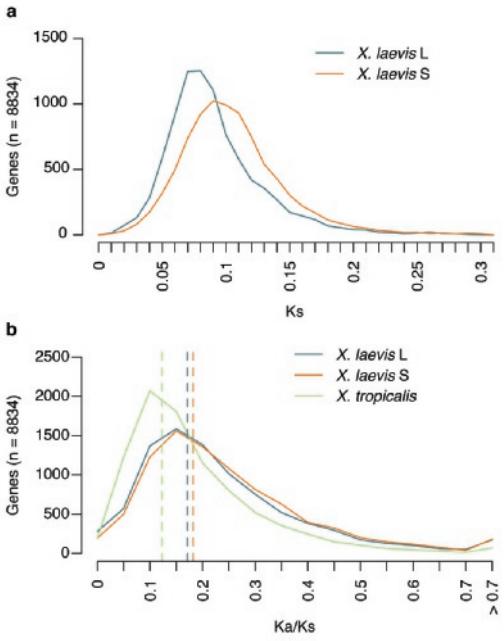


Figure 4. The S subgenome has a higher mutation rate than L. Only genes which none of the L or S copies fall into the pseudogene category are considered. **(a)** Ks distribution per subgenome in *X. laevis*. **(b)** Ka/Ks in *X. laevis* and *X. tropicalis*.

Given that the hybridization event occurred 17 Mya [8], the higher SNP density in S relative to L (Additional file 4) cannot be a relic from the time before the hybridization, (Additional file 5) and it suggests that the relatively high rate of genome degradation in S continues to this day. To examine the continuity of this genome degradation we dated unitary pseudogenes [21] caused by point mutations and / or deletion-related events (Fig. 5a). We distinguish four, non-exclusive types of pseudogenes: genes that contain a premature stop codon, genes of which the coding sequence is at least 50% shorter than their homeolog and their ortholog in *X. tropicalis*, genes that have lost at least the 75% of their promoter relative to their homeologs that do have a promoter decorated with H3K4me3 in embryos, and genes that contain a frameshift. We furthermore required for each class that the pseudogene candidate is expressed at least tenfold lower than its homeolog. In all cases, we do observe that the rate of pseudogenization has increased dramatically around 18 Mya, i.e. close to the inferred date of the hybridization, and that that rate is ~2.3-fold higher in S than in L (Fig. 5a). Furthermore, this rate continues to be high until this day for every class considered (Fig. 5b). We obtained very similar results when we included one-to-one orthologs from additional species in the dating of the pseudogenes and bootstrapped the results per gene to obtain confidence intervals (Methods, Bootstrapping pseudogene dates) (Additional file 3: Figure S5). When we separate the pseudogenes into

non-overlapping classes we observe that deletions are a prevalent cause of pseudogenization (39% and 44% on resp. L and S), and, as expected, the older pseudogenes are affected by more than one type of damage (Additional file 3: Figure S6). Pseudogenization after genome duplication has been observed to affect certain classes of protein functions more than others, with metabolic functions often being the first ones to be lost relative to regulatory proteins [6]. Indeed, when we date the loss of genes in the function categories associated with the loss, we find an overrepresentation of various metabolic processes, with the pseudogenes belonging to those categories dating often shortly after the WGD event (Additional file 3: Figure S7). We found no evidence for the preferential loss of complete complexes rather than partial complexes, e.g. for dimers the fraction of cases where of both genes only a single copy was left (17.6%), was not higher than the expected percentage if we assumed the losses of the genes from complexes to be independent from each other (18.0%) (Methods). To test for the influence of a potential dosage effect on gene loss, we compared the predicted genome-wide haploinsufficiency score (GHIS) [22] of the human ortholog of *X. laevis* homoeolog and singleton genes (Additional file 3: Figure S8). Singletons indeed have a significantly lower GHIS score than homeologs ($p = 1.1e-17$; Mann-Whitney U test), although the difference is minor (3.0%).

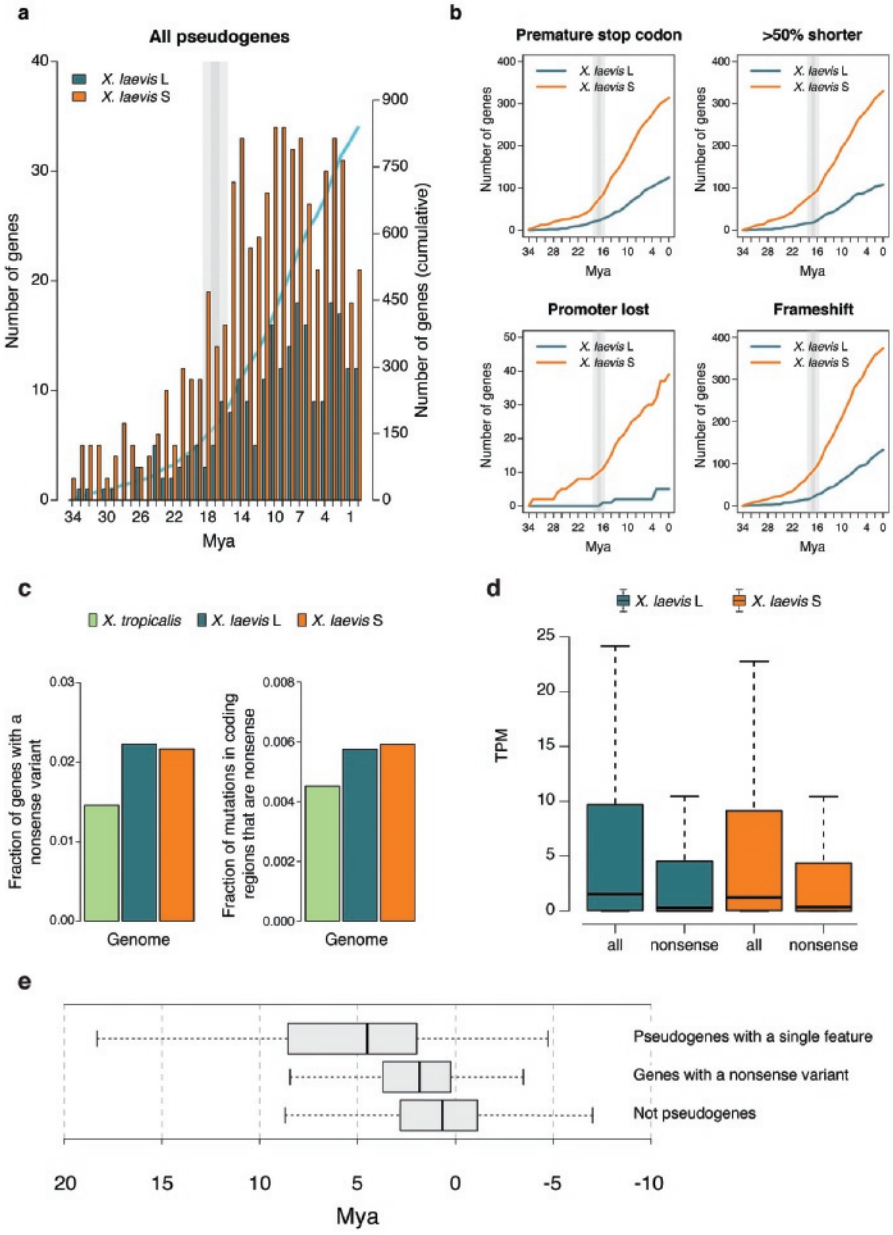


Figure 5. Pseudogenization rate has increased after hybridization (a) Number of likely pseudogenes (i.e., genes having one or more pseudogene feature and no expression while their homeolog is expressed) binned by predicted date of pseudogenization event. **(b)** Pseudogenes with different (non-exclusive) pseudogene features and their sum over the years. **(c left)** Fraction of genes that have a nonsense variant in the population. **(c right)** Fraction of mutations in coding regions that introduce a

premature stop codon. **(d)** Expression of genes with and without a nonsense variant present in the population. **(e)** Distribution of predicted pseudogenization time (including one-to-one orthologs of human, mouse, and chicken) for genes with a single pseudogene feature and a tenfold lower expression than the homeolog (top), for genes with a nonsense variant present in the population of *X. laevis* (middle) and for genes that do not present any feature for pseudogenization and whose expression is less than twofold different between homeologs (bottom).

To find independent evidence that the rate of pseudogenization in *X. laevis* remains high until the present we examined genes that appeared to be polymorphic with respect to their pseudogene state: i.e. we searched for protein truncating variants (PTVs) (variants which potentially disrupt protein-coding genes) in the progeny of two of our outbred genomes (Methods: SNP calling) relative to the published inbred genome [8]. Among all possible PTVs, we limited the analysis to SNPs that introduce a premature stop codon (nonsense mutations), as they can be called relatively reliably [23]. As a reference, we compared the nonsense SNP density with the one we measured in *X. tropicalis* using the same type of data and settings to call the SNPs: i.e. the progeny of two outbred genomes. In the 23,667 annotated genes in L and 16,939 in S we detect 528 (2.23%) and 367 (2.17%) genes with at least one loss of function variant. In contrast, in the 26,550 genes of *X. tropicalis* we detect only 388 (1.46%) loss of function variants (Fig. 5c, left). When normalizing the nonsense variants by the total number of SNPs in coding regions per (sub) genome, the fraction of premature stop variants in S (5.9e-3) is slightly higher than that in L (5.7e-3) while both are substantially and significantly higher than in *X. tropicalis* (4.5e-3; $p < 0.001$ for both comparisons; Chi-squared test; Fig. 5c, right). To substantiate that the selected PTVs are indeed hallmarks of incipient pseudogenes, we compared their expression with the expression of the other genes in their respective (sub)genome and found that genes with a SNP introducing a premature stop codon have a significantly lower expression (Fig. 5d). Second, we used the equation for dating of unitary pseudogenes to estimate the time of loss of selection in the PTV containing genes. We found that genes with this type of variants present in the population show evidence of loss of selection when compared to the set of genes that are not pseudogenes ($p = 1e-5$; Student's t-test; Fig. 5e), and that this loss of selection is more recent than for pseudogenes with only a single feature for pseudogenization that is fixed in the population ($p = 5.6e-7$; Student's t-test; Fig. 5e). That we find a higher level of SNPs in S than in L cannot be a relic from the time before the hybridization in which the S species may have had a higher SNP density than L, given that the hybridization occurred 17Mya (Supplemental note). Altogether, these results suggest that, in addition to deletions, a higher mutation rate and a more relaxed selection pressure in S has contributed to the differences that the subgenomes present nowadays, including differential gene loss. This gene loss continues to be at a higher rate than in a closely related diploid species.

Transposons have contributed subgenome-specific enhancer elements

The results described above document the pervasive loss and ongoing decay of coding and regulatory sequences after interspecific hybridization genome duplication. We next asked to

what extent regulatory innovations have contributed to genomic evolution of this species. At many loci, the profile of p300 recruitment is remarkably different between L and S loci, with differences in both p300 peak intensity and number of peak regions across homeologous loci, for example in the *slc2a2* locus (Fig. 6a). We identified 2,451 subgenome-specific p300 peaks lacking any conservation with either the other subgenome or *X. tropicalis* (colloquially referred to as 'new' enhancers). There are similar numbers of these non-conserved subgenome-specific p300-bound elements in the L subgenome (1,214) and the S subgenome (1,237).

Because new sequences can be acquired by transposition, we examined the overlap of subgenome-specific enhancers with annotated repeats and found that 87% (2,143 of 2,451; overlap >50%) are associated with annotated repeats, compared to 24% (5,557 of 23,017) of all enhancers ($p < 1e\text{-}308$; hypergeometric test). Three repeats (designated REM1, Kolobok-T2 and family-131) were particularly enriched; individually they overlap with 37-53% of the subgenome-specific p300 peaks, compared to 3-9% at other p300 peaks (Fig. 6b). Together these three annotations account for 1,338 (54%) of new enhancers, 862 of which have all three annotations overlapping at the same location. They form a 650-bp sequence with an almost perfect 195 bp terminal inverted repeat (TIR), the most terminal 65 bp of which shows 83-90% similarity with the TIRs of a Kolobok-family DNA transposon present in *X. tropicalis* (Additional file 3: Figure S9). This specific Kolobok DNA transposon carries the REM1 interspersed repeat and is present almost exclusively in *X. laevis* (8,833 and 8,802 copies in resp. L and S, respectively, vs. four copies in *X. tropicalis*), suggesting that it is a relatively young TE that proliferated after the split with *X. tropicalis*. It carries several transcription factor (TF) motifs, including the Eomes T-box motif and the Six3/Six6 homeobox motif (Fig. 6c).

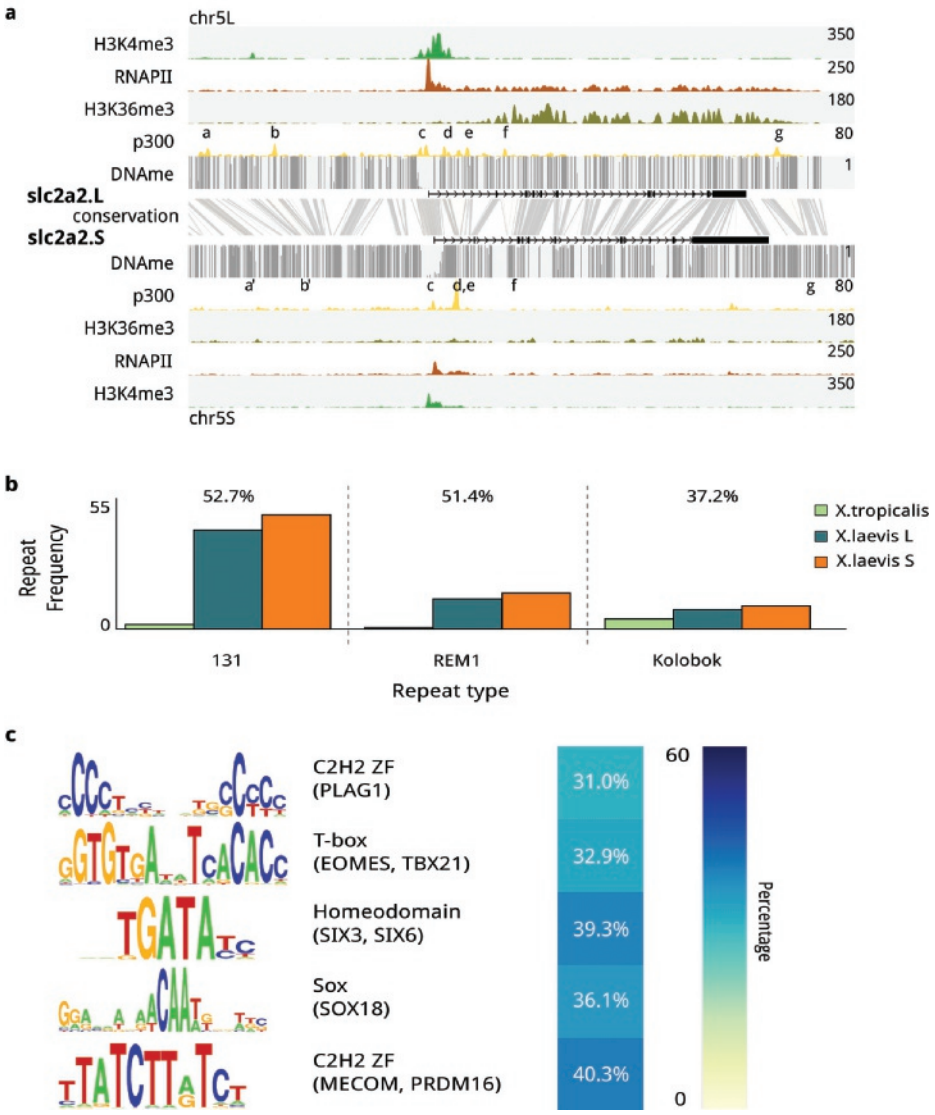


Figure 6. Sub-genome-specific recruitment of p300 is associated with TEs. Subgenome-specific p300 peaks are enriched for TEs carrying transcription factor (TF) motifs active in early development. **(a)** Differential regulation of the *slc2a2* homeologs at stage 10.5. Shown are the genomic profiles of H3K4me3 (green), RNA Polymerase II (RNAPII; purple), H3K36me3 (blue), and p300 (yellow) ChIP-seq tracks, as well as DNA methylation levels determined by WGBS (gray). The top panel shows *slc2a2.L*, which is highly expressed, as evidenced by RNAPII and H3K36me3, and has a number of active enhancers (a-g), while *slc2a2.S*, shown in the bottom panel, is expressed at a lower rate. The conservation between the L and S genomic sequence is shown in gray between the panels. Differential enhancers between L and S are highlighted in yellow, which illustrates lost enhancer function (a,b), conserved enhancer function (c-e), and deleted enhancers (f,g). **(b)** Subgenome-specific p300 peaks are associated with DNA transposon

repeats (threshold p ≤ 10e-4, twofold enrichment compared to all *X. laevis* peaks, and present at least in 15% of the peaks). The barplots show the frequency of occurrence of each of the three repeat types per megabase in the three (sub)genomes. Over the bars is represented the percentage of subgenome-specific peaks overlapping with the corresponding repeat. **(c)** TFs found to be enriched in the subgenome-specific p300 peaks (threshold p ≤ 10e-4, threefold enrichment compared to all *X. laevis* peaks, and present at least in 20% of the peaks).

We examined the correlation of the new Kolobok enhancers with gene expression and found that genes with a transcription start site within 5kb of these subgenome-specific Kolobok enhancers are more highly expressed than other genes in that subgenome (p = 1e-4 for L and p = 8e-5; Mann-Whitney U test) (Additional file 3: Figure S10), suggesting that the new enhancers are inserted close to active genes and/or promote the expression of these genes.

Regulatory remodeling by transposons in X. tropicalis × *X. laevis* hybrids

The gene expression (Fig. 2) and p300 recruitment (Fig. 6) differences between the L and S subgenomes may have been caused by regulatory incompatibilities affecting enhancer activity or DNA methylation, which could act immediately upon interspecific hybridization. Alternatively, these differences may represent the long-term effects of genomic co-evolution of the two subgenomes. To examine whether the differences between the two subgenomes were caused by the hybridization event itself, we determined the immediate effect of hybridization on DNA methylation and the patterns of H3K4me3 and p300 enrichment at regulatory regions. We generated embryos obtained by fertilization of *X. laevis* eggs (LE) with *X. tropicalis* sperm (TS). The resulting LETS hybrid embryos were compared to normal *laevis* (LELS) and *tropicalis* (TETS) embryos. The reverse hybrid (TELS) was not viable, as previously described [24].

To examine the early potential changes in DNA methylation, we performed WGBS on the DNA of LETS, LELS, and TETS embryos. The overall methylation in hybrid and normal embryos is almost identical at 92%. We identified a total of 709 differentially methylated regions (DMR) (false discovery rate [FDR]=0.05); 181 and 72 hypermethylated and 384 and 72 hypomethylated regions in respectively the *X. laevis* and *X. tropicalis* genomes. This reflects both gain and loss of DNA methylation in the sub-genomes of LETS hybrid embryos (Fig. 7f, g). There is no evidence in the underlying DNA sequence signatures for these regions being related to gene-regulatory regions (Additional file 3: Figure S11a-d). They are also not in close proximity of genes and may represent regions with inherently unstable DNA methylation. The global pattern of H3K4 trimethylation at promoters is also quite similar in hybrids and normal embryos; less than 10 peaks changed in hybrid embryos relative to normal embryos (Additional file 3: Figure S11e).

Recruitment sites of p300 however, are specifically gained and lost at several subsets of *X. tropicalis* genomic loci in hybrid embryos (Fig. 7a); 629 p300 recruitment sites were gained (a 2.6% increase relative to normal *X. tropicalis* embryos), whereas just 67 p300-bound regions

were lost (adjusted p value cutoff 1e-5). In the *X. laevis* part of the hybrid genome, none were lost or gained (Fig. 7a), indicating that the changes in the hybrid are biased towards the paternal *tropicalis* genome. To assess the epigenetic state of the gained and lost p300-binding regions, we used our epigenome reference maps of histone modifications in *X. tropicalis* [16]. Among all the marks tested, only H3K9me3 was significantly enriched, specifically at sites of gained p300 recruitment (Fig. 7b), suggesting that these regions are heterochromatic in normal (TETS) embryos but can recruit the p300 co-activator in LETS hybrid embryos.

While examining the p300 hybrid-specific recruitment sites, we noticed that transposable elements were present at many locations (Fig. 7c, d); 82% of the hybrid-specific p300 peaks overlapped more than 50% with annotated repeats. We therefore examined the occurrence of specific repeats at gained p300 sites, and found that three repeat annotations (family-451, 203 and 189) were strongly enriched (p = 1e-5; hypergeometric test), each accounting for 20-37% of all newly gained p300 peaks, whereas they only overlap with <1% of other p300 peaks (Fig. 7c, lower panel). The three repeat annotations strongly co-occur and form a 1.3 kb sequence with a 200 bp imperfect TIR, which shows ~80% similarity with those of known PiggyBac-N2A DNA transposons (Additional file 3: Figure S12). We recently found that DNA transposons that are heterochromatinized by H3K9me3 in *X. tropicalis* embryos are relatively young relative to other transposable elements [25]. Indeed, the piggyBac DNA transposons that gain p300 binding in hybrids are much less abundant in *X. laevis* than in *X. tropicalis*, suggesting that these relatively young transposons get derepressed in the *X. laevis* egg which has had little prior exposure to this transposon. These elements also carry transcription factor binding sites. Nine motifs are enriched (p = 1e-5; hypergeometric test) and are present in 10-35% of gained p300 recruitment sites, compared to a 1-3% prevalence of these motifs in other p300 peaks (Fig. 7e). These DNA binding motifs represent binding sites of Homeodomain and T-box binding factors, which are abundantly expressed during early embryogenesis.

These results document DNA transposon-associated p300 recruitment and DNA methylation instability in experimental interspecific hybrids.

DISCUSSION

The genomes of the parental *Xenopus* species that gave rise to *X. laevis* through interspecific hybridization have remarkably been maintained as separate and recognizable subgenomes propagated on different sets of chromosomes [8]. These clearly distinguishable subgenomes allow detailed analyses of the patterns of (epi)genomic loss and regulatory remodeling.

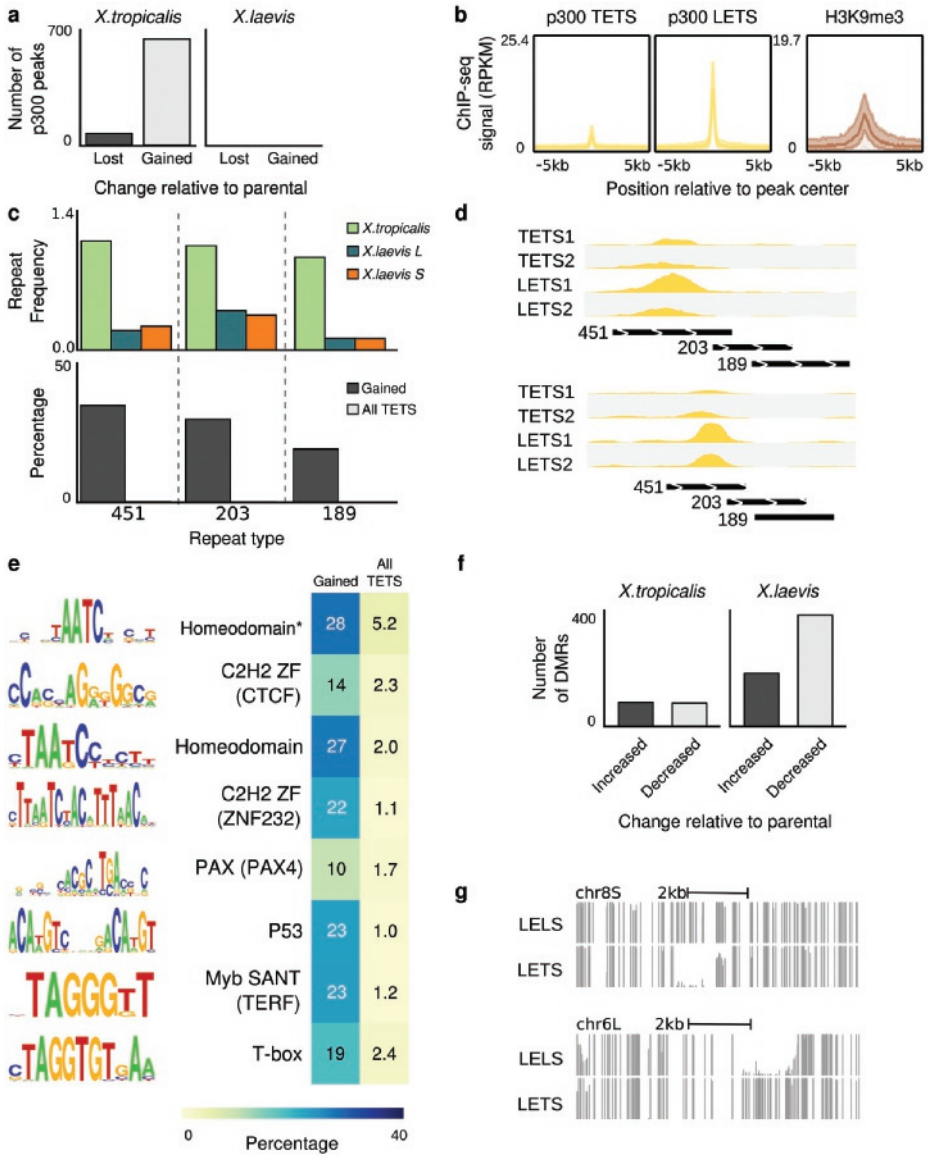


Figure 7. (a) Changes in p300 recruitment in LETS hybrids. In the *X. tropicalis* genome there are new hybridization-induced peaks as well as peaks that disappeared after hybridization. In the *X. laevis* genome there are no changes. **(b)** Newly introduced peaks appear to be repressed by H3K9me3 in *X. tropicalis* embryos. **(c bottom)** A significant number of hybrid-specific peaks are associated with DNA transposon repeats (threshold $p \leq 10e\text{-}6$, > 20 times fold enrichment compared to all *X. tropicalis* peaks and present at least in 10% of the peaks). **(c top)** The bar plots show the frequency of occurrence of Motif:lcl|rnd-1_family-451_DNA, Motif:rnd-1_family-203 and Motif:lcl|rnd-1_family-189_DNA_PiggyBac repeats per megabase in the three (sub)genomes. Those repeats are *X. tropicalis* specific, as they occur more often compared

to *X. laevis* genomes. **(d)** Profiles of *X. tropicalis* embryos p300 and LETS hybrid p300 in *X. tropicalis* hybridization-induced peaks loci. New peaks overlap with DNA transposon repeats. **(e)** Newly introduced peaks found to be enriched in TF DNA binding sites (threshold p ≤ 10e-6, fivefold enrichment compared to all *X. tropicalis* peaks, and present at least in 10% of the peaks). The TFs that can bind these motifs include Homeobox factors, C2H2 Zinc finger proteins (CTCF, ZNF232), PAX4, TERF, and T-box factors. The AATC motif, marked by an asterisk, is annotated in TRANSFAC as a GATA1 motif, but closely resembles a Paired Homeobox consensus motif. **(f)** DMRs in hybrid embryos. **(g)** DNA methylation profiles showing the DNA methylation instability in LETS hybrids.

The loss of genes, regulatory elements and genomic sequence is caused predominantly by deletions and mutations in both subgenomes, which erode the S subgenome more strongly than the L subgenome. Such biased loss of genes has been observed in polyploid plant species and has been suggested to be a general result of allo-polyploidisation, in contrast to auto-polyploidies where the subgenomes are indistinguishable and degrade at a similar rate [9]. As to why one particular subgenome erodes more quickly than another, one hypothesis is that interspecific hybridization generates a crisis, referred to as 'genomic shock', for example by transposon reactivation on one of the subgenomes which can disrupt coding sequences [26]. Consistent with this possibility is the proliferation of S-specific Mariner DNA transposons in *X. laevis* at the time of hybridization [8]. Also consistent with transposon reactivation are our results from artificial *X. tropicalis* × *X. laevis* hybrids (LETS, *X. laevis* eggs, *X. tropicalis* sperm), in which a set of *X. tropicalis*-specific DNA transposons recruits the p300 co-activator in the hybrid, whereas normally they are repressed by H3K9me3. Relatively young DNA transposons are heterochromatinized with H3K9me3 [25], but when introduced into eggs that have been little exposed to these transposons these mechanisms may fail. We have not been able to detect transposon expansion in the short time of *Xenopus* hybrid embryogenesis (data not shown), but together the observations suggest that transposon reactivation can contribute to genomic perturbations in hybrids. Similarly, in the Atlantic salmon, which has undergone several (320 Mya, 80 Mya) whole genome duplications, transposon expansion has been associated with the whole genome duplication event and with chromosome rearrangements [6].

In contrast to these short-term effects of hybridization, our analyses indicate that new pseudogenes continue to arise, both by mutations that cause premature stop codons, and by deletions that truncate the coding region or delete intergenic or promoter regulatory sequences. An elevated rate of pseudogene formation is observed on both the L and S subgenomes since the time of hybridization (~17 Mya, cf. Fig. 5) up to the present day, suggesting genome erosion is a continuous process that has been and still is higher on S compared to L. Consistent with this result is a mildly elevated level of SNPs observed in S relative to L (Fig. 4; Additional file 4). The cause of the higher mutation rate of the S subgenome is unknown. The local mutation rate has been shown to correlate with replication timing [27] and it is possible that there are subtle but consistent differences in replication timing between the two subgenomes. It can

also be due to differences in background selection [28], in which selection against non-neutral variants would also reduce neutral variation in their vicinity.

All in all, the higher level of genome degradation in S relative to L appears to be the result of a slightly higher mutation rate and a considerably higher deletion rate in S, combined with less selection against the loss of (epi)genetic elements in S than in L. The higher deletion and mutation rates are supported by higher numbers of deletions and SNPs in regions that appear not to be under selection: intergenic regions, introns and redundant coding positions. Reduced selection against the loss of genetic elements from S relative to L is supported by a larger difference in the loss of p300 peaks and promoters relative to the background in the L subgenome than in the S subgenome and a slightly but significantly lower Ka/Ks ratio in the L subgenome relative to the S subgenome.

The deletions bear the hallmarks of NAHR [29]; the retained regions in the other subgenome are enriched for ancient repeats and the sequence similarity between the flanks of the region is higher than expected by chance. The S chromosomes have also experienced significantly more rearrangements including inversions [8]. Normally, in meiotic recombination double strand breaks are fixed using allelic sequences. In the absence of proper chromosome pairing, other non-allelic homologous sequences, for example repeats in the same chromosome, are used for double-strand break repair, leading to deletions and inversions [29]. Interestingly, Prdm9, a fast-evolving mammalian DNA-binding protein involved in meiotic chromosome pairing and recombination hotspot selection, has been implicated in hybrid sterility in mouse [30], [31]. There is no known one-to-one ortholog of Prdm9 in *Xenopus* and the L and S subgenome-encoded proteins involved in meiotic double strand break repair are also not fully known, but it is conceivable that their skewed expression or activity is involved in subgenome-biased NAHR.

The results reported here identify a major role for repetitive elements in subgenome bias, gene loss and regulatory remodeling. Not only is sequence loss by NAHR linked to repeats, subgenome-specific acquisition of enhancer elements is also overwhelmingly associated with TEs. Moreover, young transposons also gain p300 recruitment in *X. tropicalis* × *X. laevis* hybrids. DNA transposons can contribute sequence variation to the genome, which can affect gene expression by changing the local chromatin state at the site of insertion, resulting in metastable epi-alleles [26]. Once a host is invaded, transposable elements usually duplicate freely before they become repressed. When introduced in relatively unexposed eggs this repression may be lost. Interestingly, transposable elements can be co-opted as enhancers for the regulation of developmental genes [32], [33]. TFs have been found to bind to TEs with open and active chromatin signatures in both human and mouse cells, but the binding patterns were largely different between the two species [34], suggesting that transposons contribute to

regulatory change during evolution. In addition to the potentially large and sudden changes in regulatory potential caused by transposition, mutational changes are known to cause TF binding sites to be lost and gained [17], [35], causing turnover and change in the regulatory landscape over longer time scales.

CONCLUSIONS

It is not known exactly how the ancient two rounds of whole genome duplications at the root of the vertebrate tree have contributed to genome evolution. Its analysis is confounded by the pervasive loss of homeologs over hundreds of millions of years and the absence of tractable subgenomes. The *X. laevis* interspecific hybridization and genome duplication event is one of the most recent vertebrate genome duplications. Excitingly, the clearly distinguishable chromosomes of different parental origins allow for reconstruction of the parental genomes. We have found evidence for a pervasive influence of repetitive elements, driving gene loss and genomic sequence loss through NAHR, in addition to remodeling of the regulatory landscape through transposon-mediated gain of coactivator recruitment. In combination with experimental interspecific hybrids, *Xenopus* can therefore be a powerful new model system to distinguish the short and long-term consequences of hybridization and to study the mechanisms of vertebrate genome evolution.

METHODS

Animal procedures

Embryos were generated using IVF (in vitro fertilization) with outbred animals, including LELS embryos (*laevis* eggs-*laevis* sperm), TETS embryos (*tropicalis* eggs-*tropicalis* sperm) and LETS embryos (*laevis* eggs-*tropicalis* sperm). *X. laevis* female frogs were injected with 500U of hCG (human chorionic gonadotropin, BREVACTID 1500 I.E) 16 hours before IVF. A *X. laevis* male was sacrificed and isolated testis was macerated in 2 mL Marc's Modified Ringer's medium (MMR) to be used immediately for fertilization. Both male and female *X. tropicalis* frogs were primed with 100 and 15U of hCG 48 hours before IVF. Five hours prior to egg laying, females were boosted with 150U of hCG. Male testis was always isolated fresh. The testis was macerated in 2 mL FCS-L15 (10% fetal calf serum-90% L15 medium) cocktail and used immediately for IVF. LETS embryos were obtained similarly using species and sex-specific hormonal stimulation as described above. Once the macerated sperm suspension was mixed vigorously over the layered eggs, they were left undisturbed for three minutes and then the Petri dish was flooded with 25% MMR for the fertilized *X. laevis* eggs (LELS and LETS) and 10% MMR was added to the fertilized *X. tropicalis* eggs (TETS). Embryos were cultured at 25°C. The jelly coats were removed 4 hpf (hours post-fertilization) using 2% cysteine in 25% MMR (pH 8.0) for LELS and LETS and using 3% cysteine in 10% MMR (pH8.0) for TETS.

ChIP-sequencing

Embryos (n = 35-90, two biological replicates for every ChIP experiment) were fixed in 1% formaldehyde for 30 minutes at Nieuwkoop-Faber stage 10.5. Embryos were washed once in 125 mM glycine / 25% Marc's Modified Ringer's medium (MMR) and twice in 25% MMR, homogenized on ice in sonication buffer (20 mM Tris•HCl, pH 8/10 mM KCl/1mM EDTA/10% glycerol/5 mM DTT/0.125% Nonidet P-40, and protease inhibitor cocktail (Roche)). Homogenized embryos were sonicated for 20 minutes using a Bioruptor sonicator (Diagenode). Sonicated extract was centrifuged at top speed in a cold table-top centrifuge and supernatants (ChIP extracts) were snap frozen in liquid nitrogen and stored at –20°C until use. Prior to assembling the ChIP reaction, the ChIP extract was diluted with IP buffer (50 mM Tris•HCl, pH 8/100 mM NaCl/2mM EDTA/1 mM DTT/1% Nonidet P-40, and protease inhibitor cocktail) and then incubated with 1–5 µg of antibody and 12.5 µl Prot A/G beads (Santa Cruz) for an overnight binding reaction on the rotating wheel in the cold room. The following antibodies were used: H3K4me3 (Abcam ab8580), H3K4me1 (Abcam ab8895), p300 (C-20, Santa Cruz sc-585), H3K36me3 (Abcam ab9050) and RNA polymerase II (Diagenode C15200004). The beads were sequentially washed, first with ChIP1 buffer (IP buffer plus 0.1% sodium deoxycholate), then ChIP2 buffer (ChIP1 buffer with 500 mM NaCl final concentration), then ChIP3 buffer (ChIP1 buffer with 250 mM LiCl), then again with ChIP1 buffer, and lastly with TE buffer (10 mM Tris,

pH 8/1 mM EDTA). The material was eluted in 1% SDS in 0.1 M sodium bicarbonate. Cross-linking was reversed by adding 16 µl of 5 M NaCl and incubating at 65°C for 4–5 hours. DNA was extracted using the Qiagen QIAquick PCR purification kit. ~ 10 ng input DNA was used for sample preparation for high-throughput sequencing on an Illumina HiSeq 2000 or NextSeq (according to manufacturer's protocol).

RNA-sequencing

For RNA-sequencing experiments total RNA was extracted from 20 Nieuwkoop-Faber stage 10.5 embryos (two biological replicates each for LELS and LETS respectively) using Trizol and Qiagen columns. 4-5 µg of total RNA was treated with DNase I on column and depleted of rRNA (ribosomal RNA) using Magnetic gold RiboZero RNA kit (Illumina) resulting in a yield of 45 - 52 ng of rRNA depleted total RNA. 2 ng of rRNA-depleted total RNA was reserved for Experion (Bio-Rad) quality assessment run for rRNA depletion and the remaining was used for first and second strand synthesis (strand-specific protocol). Total yield of dscDNA was between 14.5-15.8 ng and out of this 1.2 - 5 ng was used for sample preparation for high high-throughput sequencing (according to manufacturer's protocol). qPCR quality controls before and after sample preparation corroborated well and relative depletion of 28S rRNA compared to control genes (eef1a1 and *gs17*) was taken as a quality assessment indicator for sequencing-grade dscDNA.

ChIP-seq and RNA-seq data analysis

ChIP-seq reads were mapped to the *X. laevis* genome (Xenla9.1) using bwa mem (version 0.7.10-r789) with default settings [36]. Duplicate reads were marked using bamUtil v1.0.2. Where applicable (H3K4me3, p300) peaks were called using MACS (version 2.1.0.20140616) [37] relative to the Input track using the options --broad -g 2.3e9 -q 0.001. --buffer-size 1000. Peaks were combined for replicates using bedtools intersect (version v.2.20.1) [38]. Figures of genomic profiles were generated using fluff v1.62 [39].

In addition to the RNA-seq triplicate produced in this study, we used the eight stage 10.5 samples from NCBI GEO series GSE56586 (GSM1430926, GSM1430927, GSM1430928, GSM1430929, GSM1430930, GSM1430931, GSM1430932, GSM1430933). RNA-seq reads were mapped to the Xenla9.1 genome with the JGI 1.8 annotation using STAR version 2.4.2a [40]. Quantification of expression levels was performed using express eXpress version 1.5.1 [41]. The mean expression level (TPM; transcript per million) per transcript was obtained by combining all replicates.

MethylC-seq for whole-genome bisulfite sequencing

Genomic DNA from *Xenopus* embryos (LELS and LETS, n = 20-50, NF stage 10.5) was extracted as described before [42] with minor modifications. Briefly, embryos were homogenized in 3 volumes STOP-buffer (15 mM EDTA, 10 mM Tris-HCl pH7.5, 1% SDS, 0.5 mg/mL proteinase K). The homogenate was incubated for 4 hours at 37 . Two phenol:chloroform:isoamyl alcohol (PCI, 25:24:1) extractions were performed by adding 1 volume of PCI, rotating for 30 minutes at RT (room temperature) and spinning for 5 minutes at 13k rpm. DNA was precipitated in 1/5 volume NH4AC 4M plus 3 volumes EtOH with an overnight incubation at 4 . Subsequently, the DNA was spun down for 20 minutes at 13k rpm in a cold centrifuge and the pellet was washed with 70% EtOH and dissolved in 100 µL of DNAse free water. To remove contaminating RNA, a 2 hours RNase A (0.01 volume of 10 mg/mL) treatment was performed at 37 . Sample was further purified with two Mg/SDS precipitations. 0.05 volumes of 10% SDS plus 0.042 volumes of MgCl2 2M was added to the sample followed by incubation on ice for 15 minutes. Subsequently, the precipitants were spun down at 4 for 5 minutes at 13k rpm. A third PCI extraction was also performed followed by only one chloroform:isoamyl alcohol (CI, 24:1) extraction. DNA was precipitated overnight at -20 in 2.5 volumes EtOH plus 1/10 volume NaOAc 3M pH 5.2. Next, the precipitated DNA was spun down for 30 minutes at 13k rpm in a cold centrifuge and the pellet was washed with 70% EtOH. The purified DNA pellet was then dissolved in 50 µL H2O.

MethylC-seq library generation was performed as described previously [43], [44]. The genomic DNA was sonicated to an average size of 200 bp, purified and end-repaired followed by the ligation of methylated Illumina TruSeq sequencing adapters. Library amplification was performed with KAPA HiFi HotStart Uracil+ DNA polymerase (Kapa Biosystems, Woburn, MA), using 6 cycles of amplification. MethylC-seq libraries were sequenced in single-end mode on the Illumina HiSeq 1500 platform. The sequenced reads in FASTQ format were mapped to the in-silico bisulfite-converted *Xenopus laevis* reference genome (Xenla9.1) using the Bowtie alignment algorithm with the following parameters: -e 120 -l 20 -n 0 as previously reported [45], [46]. Differentially methylated regions were called using the methylpy pipeline, as described before [46], with FDR < 0.05 and the difference in fraction methylated larger than or equal to 0.4. To estimate the bisulfite non-conversion frequency, the frequency of all cytosine base-calls at reference cytosine positions in the lambda genome (unmethylated spike in control) was normalized by the total number of base-calls at reference cytosine positions in the lambda genome. See below for sequencing and conversion statistics.

DNA-methylation free (hypo-methylated) regions were detected using the hmr tool from MethPipe version 3.0.0 (http://smithlabresearch.org/software/methpipe/) [47]

Active transcription

To consider a region as actively transcribed, we measured the H3K36me3 and RNAPII marks (as RPKM) of 200.000 random regions in *X. laevis* to define background levels. Regions with active transcription are those with at least the average of the measures plus two standard deviations, for both signals independently.

Whole-genome alignment

Genome alignment of *X. tropicalis* and *X. laevis* was performed using progressiveCactus version 0.0 (https://github.com/glennhickey/progressiveCactus) [39], [40] with the default parameters. *X. tropicalis*, *X. laevis* L and S were treated as separate genomes and were aligned using (Xla.v91.L:0.2,Xla.v91.S:0.2):0.4,xt9:0.6) Newick format phylogenetic tree. In order to reduce computational time alignment was done per-chromosome, with homeologous chromosomes aligned to each other.

Calling deletions

A set of high-confidence deleted regions was obtained using the progressiveCactus alignment. We extracted all regions from the *X. laevis* genome that reciprocally aligned either *X. tropicalis* and/or to the other subgenome. We then selected all regions that reciprocally aligned to *X. tropicalis* but not to the other *X. laevis* subgenome. We merged all regions within 10 bp and removed those that overlapped for more than 25% of their length with gaps. As a final filtering step, we required a sequence that reciprocally aligned to the other subgenome in both 500 bp flanks of the putative deletion. Finally, the size of the region between the two aligned flanks should be at most 4kb and at least 3 times shorter than the size of the region in the subgenome where the sequence was not deleted.

SNP calling

SNPs were called using the GATK pipeline (version 3.4-46-gbc02625 [48]) on basis of the best practices workflow [49], [50]. As input we used a high-coverage ChIP-input track from a clutch of wild-type embryos compared the reference J-strain genome. The HaplotypeCaller tool was used to call SNPs. All putative SNPs were subsequently filtered with the VariantFiltration tool. The filterExpression was set to "QD < 2 || FS > 60.0 || MQ < 35.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" for *X. tropicalis*. For *X. laevis* the same settings were used, except for MQ, which was set to "MQ < 40". SNPs passing the filter were required to have at least ten-fold coverage with at least four observations of the alternative allele. The SNP coverage was calculated relative to the sequence regions where SNPs could be called given the minimum required coverage, as determined by the CallableLoci tool from the GATK pipeline.

Search and alignment of orthologs and evolution rates

Orthologs of *X. tropicalis* were searched in the genome of *X. laevis* with the cdna2genome tool from Exonerate [51]. From 14500 sequences submitted, 14276 were successfully scanned. From those, 10935 found a match in both subgenomes, leaving 3343 sequences that did not return any sequence from either L or S subgenomes or both. Among the sequences with a match in both subgenomes, those having no synteny (939) were discarded because they were potential wrong matches in closely related gene families.

Once we had our three sequences per gene (9996), we aligned them using MACSE [52], which allows frameshifts and premature stop codons, with the following parameters: gap creation -18, gap extension -8, frameshift creation -28, premature stop codon -50. 10 sequences were discarded in this step.

In order to obtain evolutionary rates of each of the three copies per gene triangle, we performed ancestral sequence reconstruction with FastML [53], which gave us the most likely sequence present at the speciation between *X. laevis* L and S ancestors. Once we obtained this crossroad sequence, we measured the amount of ratio of nonsynonymous mutations per nonsynonymous sites versus synonymous mutations per synonymous sites (i.e., Ka/Ks ratio) using the seqinR package [54].

Pseudogene dating

Similar to Zhang et al. [21] we related the excess of nonsynonymous mutations to the evolving rate average of the gene to date the approximate time when the copy lost constraint on its sequence.

Bootstrapping pseudogene dates

We took the pseudogene candidates and retrieved their annotated 1 to 1 orthologs in human, mouse and chicken through Ensembl. We then aligned them using MACSE [52] with default parameters, considering the pseudogene as a “less reliable” sequence. After this, we reconstructed the ancestral sequence with FastML [53] and then measured the Ka/Ks ratio using the seqinR package [54].

In order to confirm the reliability of these results, we bootstrapped the alignments 1000 times each and measured the Ka/Ks ratios of all of them. Briefly, we cut up the alignments in codons and we built an artificial alignment of the same length of the original protein by randomly adding (with replacement) aligned codons found in the original alignment.

Quantification of genomic losses per genomic region

Using the deletions track generated through the deletions call step (see Methods: Calling deletions), we quantified the amount of DNA lost per genomic region by measuring the overlap between both coordinates. To do so, we used the R packages rtracklayer [55] and GenomicRanges [56]. To compare the observed distribution of deletions to the expected distribution, we performed 1000 genomic randomizations of the deletions, keeping features on the same chromosome, using bedtools shuffle [38] with the -chrom argument. P-values for enrichment or depletion of overlap with specific features were calculated based on the z-score obtained from the 1000 randomizations. P-values for differences in observed/expected rate between L and S chromosomes were calculated using the Mann-Whitney U test. All P-values were adjusted for multiple testing using the Benjamini-Hochberg approach.

Gene Ontology term enrichment analysis

Term enrichment analysis was performed using PANTHER [57]. Briefly, we used *X. tropicalis* orthologs names of the pseudogenes discussed in section 6 and we compared it to the list of genes in *X. tropicalis* that successfully returned syntenic orthologs in *X. laevis* (see Methods: Search and alignment of orthologs and evolution rates).

Quantification of preferential loss of complete protein complexes

We took the hetero-dimers from the human protein complex CORUM database [58] and examined the extent to, when completely represented in the *X. laevis* genome (357 complexes), both genes were present on both genomes (170 complexes), only one gene was present on both genomes (124 complexes), or both genes were present on only a single genome (63 complexes). Also, extending the analysis to trimers did not show an over representation of completely lost complexes.

DECLARATIONS

Acknowledgements

The authors thank Ulrike J. Jacobi and Kees-Jan François for valuable contributions in an early phase of the work and Emese Gazdag for genomic DNA.

Funding

This work has been supported by the US National Institutes of Health (NICHD, grant R01HD069344). Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. DME and MAH were supported by the Virgo consortium, funded by the Dutch government (FES0908). RG was supported by an HFSP long term fellowship LT 0004252014-L. RH was supported by R35 GM118183. SJvH is supported by the Netherlands

Organization for Scientific research (NWO-ALW, grant 863.12.002). OB is supported by an Australian Research Council Discovery Early Career Researcher Award - DECRA (DE140101962).

Availability of data and materials

The data have been deposited in NCBI's Gene Expression Omnibus [35] and are accessible through GEO Series accession numbers GSE76059 (*X. laevis* ChIP-seq), GSE92382 (genomic DNA; *X. laevis* RNA-seq; *X. tropicalis* × *X. laevis* ChIP-seq), GSE90898 (*X. tropicalis* × *X. laevis* whole-genome bisulfite sequencing) and GSE67974 (*X. tropicalis* ChIP-seq).

Authors' contributions

ChIP-seq, RNA-seq data generation and experimental design was performed by SSP with help from IvK, RG and RH. GJCV, SJvH and MAH designed the study. DME and GG were involved in analysis design. Bisulphite sample generation and sequencing was done by SSP, IvK and OB, RL. OB and GG performed analysis of differentially methylated regions. Genome alignment and hybrid analysis was performed by GG. Analysis of deleted regions and SNPs was performed by SJvH and DME. DME also performed analysis of mutation rates and pseudogenes. DME, SSP, GG, MAH, SJvH and GJCV wrote the paper. DME, SSP, GG contributed equally to the study. All authors discussed the results and commented on the manuscript.

SUPPLEMENTARY FIGURES

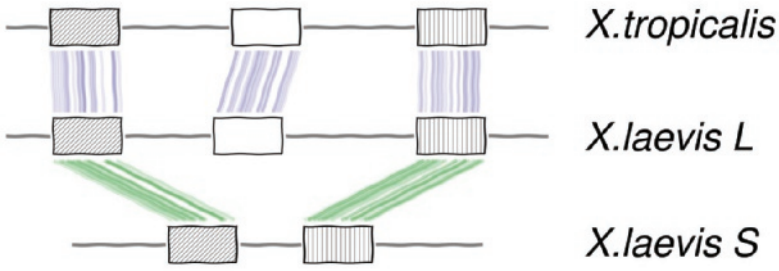


Figure S1. Strategy for calling deletions based on blocks of sequence conserved between one *X. laevis* subgenome and *X. tropicalis*, but lost from the other subgenome.

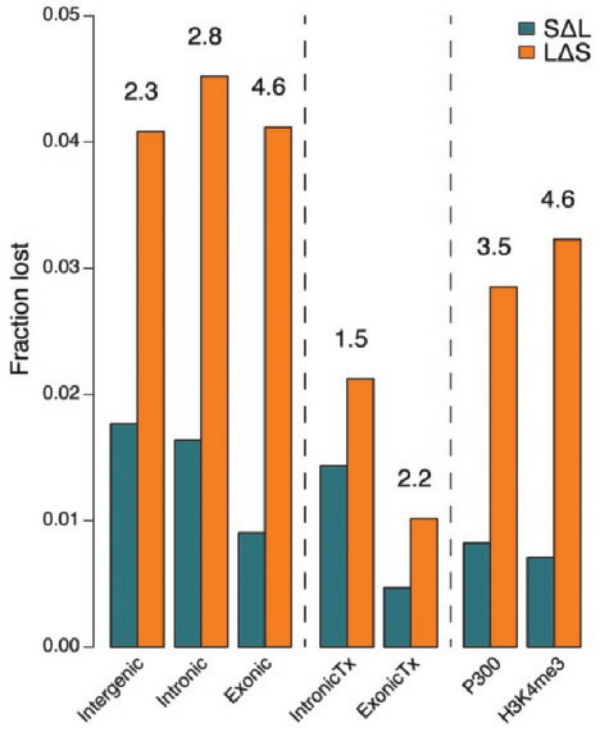


Figure S2. Fraction of genomic regions lost by deletions. Numbers on top of the bars represent the ratio of the fraction lost in S relative to the one lost in L. Intergenic: 1kb distance from a gene. Intronic: introns. Exonic: UTRs + CDS. IntronicTx: introns from genes actively transcribed. ExonicTx: Exons from genes actively transcribed. p300: genomic fragments having a p300 peak. H3K4me3: genomic fragments having a H3K4me3 peak.

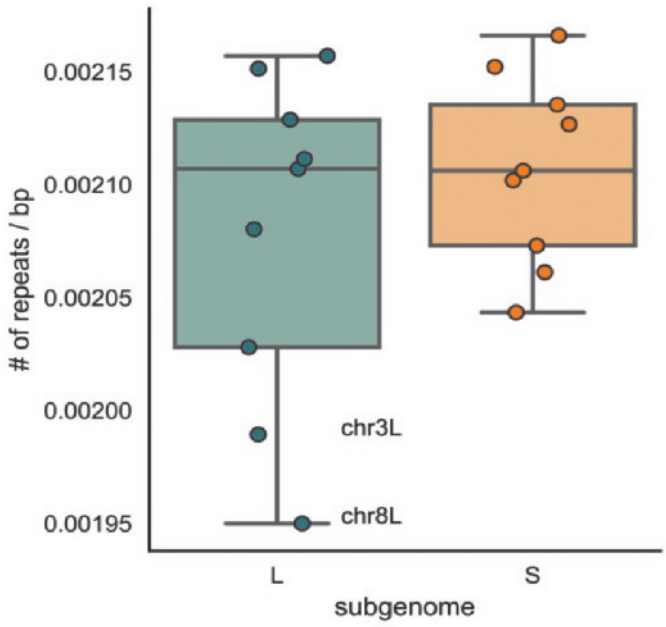


Figure S3. Number of repeats on L and S chromosomes.

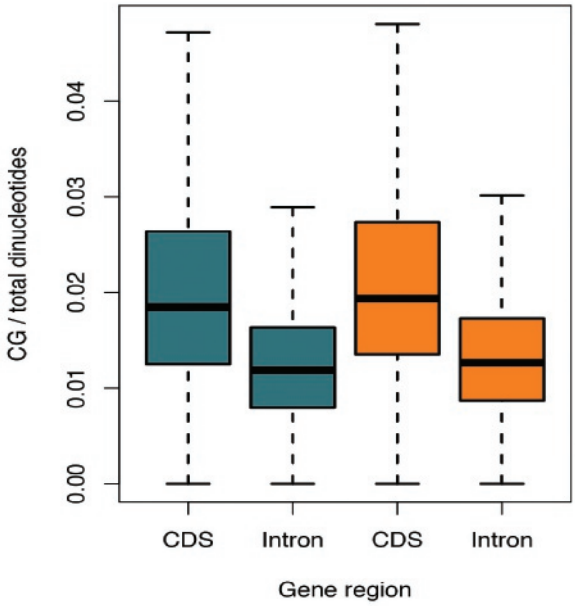


Figure S4. CpG density is in CDS and in introns for both L and S subgenomes.

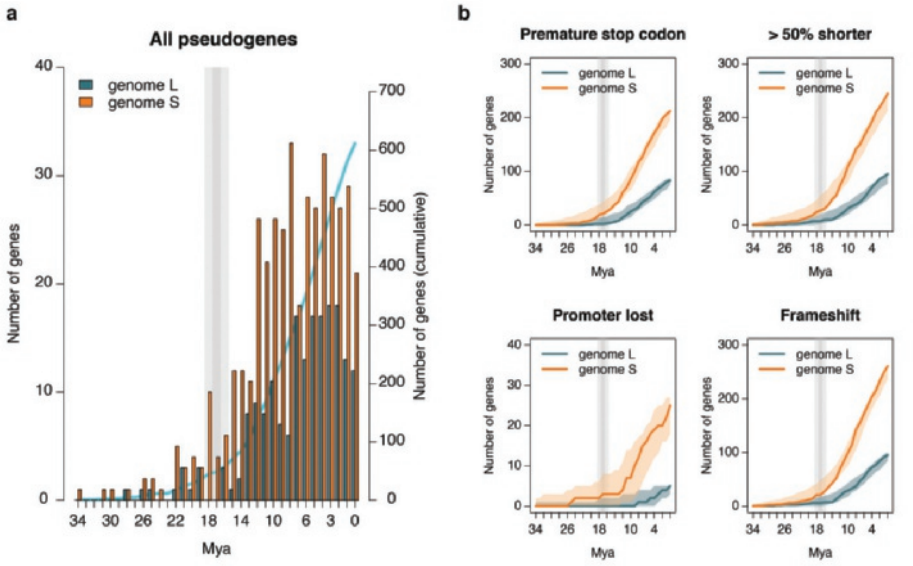


Figure S5. Dating of pseudogenes using extra species. (a) Number of likely pseudogenes (i.e., a gene presenting one or more pseudogene features and 10 times less expression than its homeolog) which have been successfully aligned to their orthologs in human, mouse and chicken, binned by predicted date of pseudogenization event. **(b)** Likely pseudogenes with different (non-exclusive) pseudogene features and their sum over the years. The shaded area depicts the upper and the lower estimates based on the results of the bootstraps.

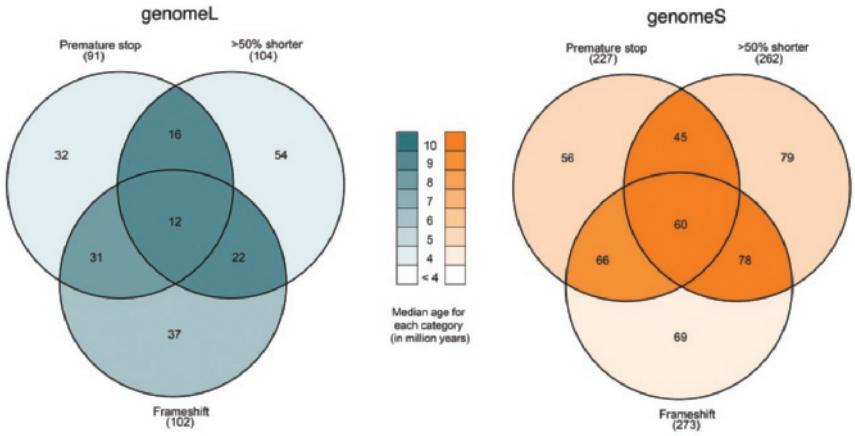


Figure S6. Median age of each category in each subgenome for pseudogenes with one-to-one orthologs in human, mouse and chicken.

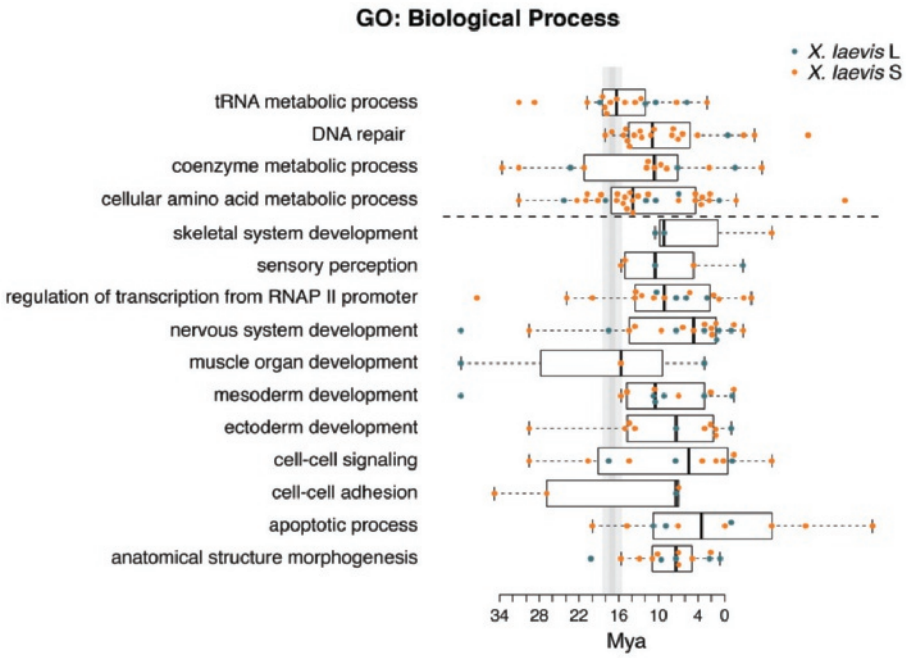


Figure S7. GO term enrichment analysis. Each dot is a pseudogene and its predicted pseudogenization time.

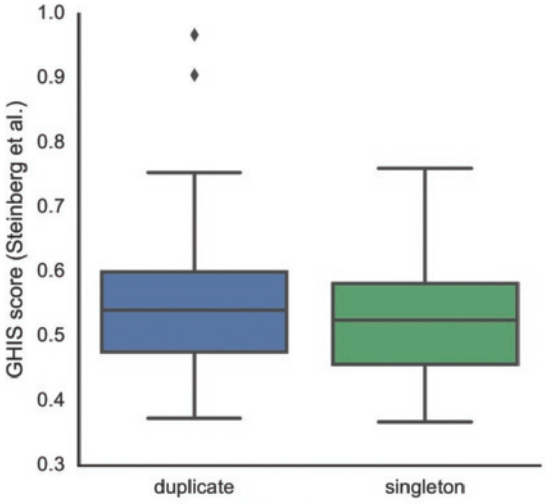
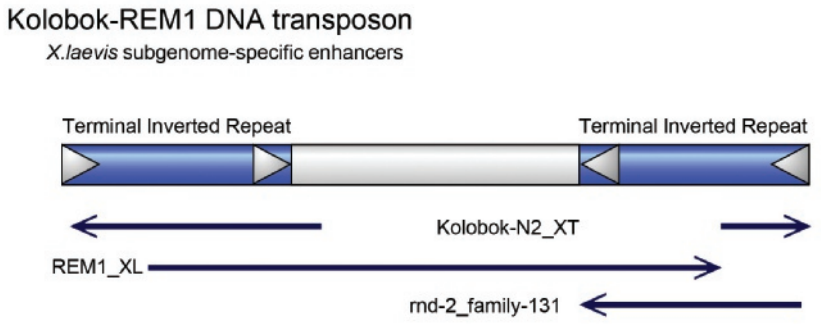


Figure S8. The distribution of genome-wide haploinsufficiency scores (GHIS;Steinberg et al.) of the human homologs of *X. laevis* genes that are present in eitehr one copy (singleton) or two (duplicate; homeologs). Genes that are retained as two copies have a significantly higher GHIS score (p=1.09e-17, Mann-Whitney U).



>Kolobok-REM1 DNA transposon (consensus of 9)
TTAAAGGGATACTGTCATGGGAAAAAAANATTTTTTTCAAAATGAATCAGTTAATAGTGCTGCTCCAGCAGAATTCTGCA
CTGAAATCCATTTCTCAAAAGAGCAAACAGATTTTTTTATATTCAATTTTGAAATCTGACATNGGGGCTAGACATATTGT
CAATTTCCCAGCTGCCCCAAGTCATGTGACTTGTGC TCTGATAAACTTCAATCACTCTTTACTGCTGTACTGCAAGTTGG
AGTGATATCACCCCCTNCCNTTTTCCCCCCCAGCAGCCAAACAAAAGAACAATGGGAAGGTAACCAGATAGCAGCTCCCT
AACACAAGATAACAGCTGCCTGGTAGATCTAAGAACAACACTCAATAGTAAAAACCCATGTCCCACTGAGACACATTCAG
TTACATTGAGAAGGAAAAACAGCAGCCTGCCAGAAAGCATTTCTCTCCTAAAGTGCAG GCACAAGTCACATGACTTGGGG
CAGCTGGGAAATTGACAAAATGTCTAGCCCCATGTCAGATTTCAAAATTGAATATAAAAAAATCTGTTTGCTCTTTTGAG
AAATGGATTTCAGTGCAGAATTCTGCTGGAGCAGCACTATTAACTGATTCATTTTGAAAAAAATATTTTTTCCCATGACA
GTATCCCTTTAA

Figure S9. Structure (top) and sequence (bottom) of *X. laevis* Kolobok-REM1 DNA transposon which can recruit the p300 co-activator. The component annotations from the repeat track are shown.

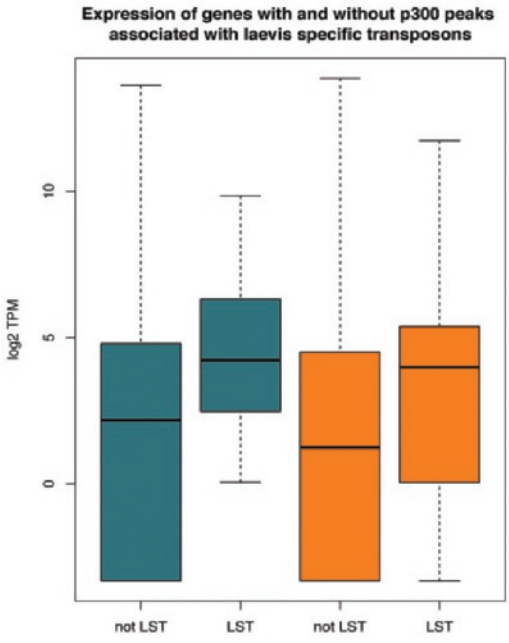


Figure S10. Expression of genes without and with new p300 peaks in laevis specific transposons.

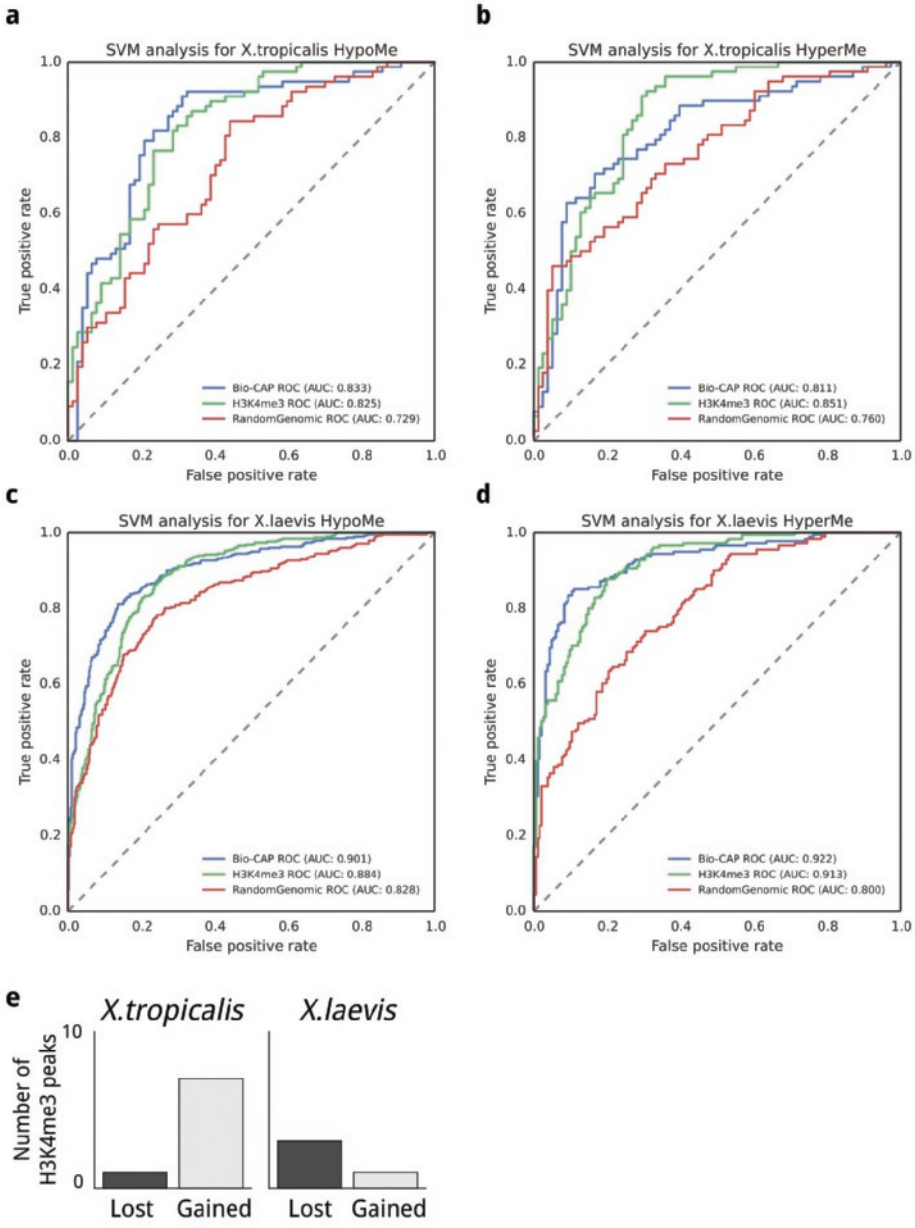
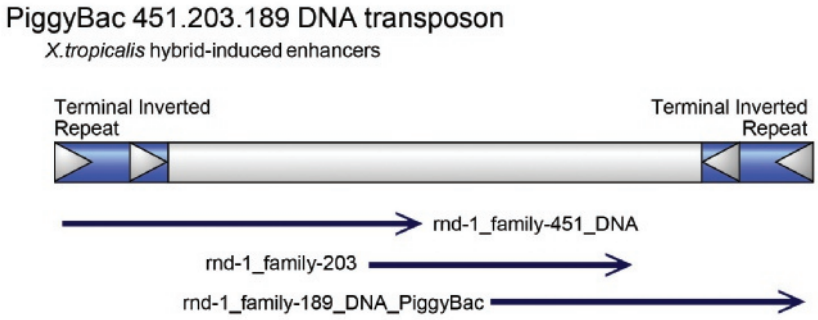


Figure S11. Analysis of differentially methylated regions (DMRs) and H3K4me3 in hybrid embryos. (a-d) Receiver-Operator Curves (ROC) of Support Vector Machines trained on DMR k-mers versus k-mers present in random genomic DNA (red), H3K4me3-positive promoter regions, and unmethylated regions (blue) profiled using Bio-CAP [55]. Areas under the curve (AUC) > 0.5 imply that the SVM distinguishes DMR sequence from other sequences in the case of lost (a) or gained (b) DNA methylation in the *X. tropicalis* subgenome of hybrid embryos, and lost (c) or gained (d) DNA methylation in the *X.*

laevis subgenome of hybrid embryos. DMR sequences appear to be different from promoters (H3K4me3 ChIP-seq peaks), unmethylated CpG islands and random genomic sequences, suggesting they represent a specific subset of genomic sequences. However, DMRs with gained DNA methylation (HyperMe) were indistinguishable from DMRs with lost DNA methylation (HypoMe; not shown). A similar result was obtained in the comparison between DMRs present in the *X. laevis* L and S genomes. This indicates that there are no specific sequence signatures distinguishing different types of DMRs (de novo methylated or demethylated, *X. tropicalis* or *X. laevis*). **(e)** Virtually no gain or loss of H3K4me3 peaks was observed in the subgenomes of LETS hybrid embryos.




```
>PiggyBac 451.203.189 DNA transposon (consensus of 9)
GGGGGAATTCACAAAAGTGGAGANAGCCCTTTTTTGGAGTAAAAGTGGTGGAAAAACTGGTGGAAAACACTTTCTCCAGA
TTCACAAACAGAAATCCGCCTAAATTCCGACTCAACCGCCACTTNTCCGTTTTTGGTGAAAGAAAGACAGTTTACAGACA
CTTTTGTGAATTTGTCGTTTAAACGACATTTATACGACGTTTTAACAACATTTTTTCCCNACATTTTCAAAATGGAGTAA
AGCTCAGTGTAACTCTGTCAGATCAATTCTAAGCTCTTCTGTTTTGTTTTGTTTCACCCAGTCTCCATTTCACACCTAAG
GTAGGGGCCCCATTGGGGGATCATTAACATATTAATGGGGATTAGCAGCCTATTGTTAGGGGACAAGTTTCTGTCTAACC
CTAGAACAATAGGTGCAAAAAGCCTCATTAACATTTTATAAACAAGTAAAACACTGATTAAANAGTTTAATTTATATCTT
ATATATAAAAAGTTTTTACCTCACATTTGCATGATTATGCTAGTTTTAGCTTAATCAATGAGGCAATAATAACATTTTGT
GTGAATATATTGTAAACATTTTTATTAAAAGGAAAAGTAAAGCCTCCCAGGCAAATTTTTAGTTTTAGCAGCAGTGCCCC
CTCTTTAATCACTTCACTGACATTTGCCCCAACTTATCTGTGCCCCCATAGCATGTCCAGAACTACAGAGCTGCTTGACA
GCATTGGTCCCNACCTGTTCCCAGCAGCCATCTTGGTGATCAGCAAACAGCACATACACACTGGCACCCAAACTGGGATA
TTGGGGTACAGGGTTGGACTGGCCCAACAGGGCACTGGATAAAAACCCAGTGGGCCCTAGTCCTGTTGGACCTCTATTCA
CCGGGCATGTTGGTGTGTGGGCCGGATTGCTGCTGNCCAGTAATGAGTAGGGGCTAGAGAAGAAGGTATTTAGACCCTTA
AATCTTGTTACAGAAGTGGAATTACACAGAAATATAGAAAACATTAGAAAGCCCAGATTGTCCTGAAAAAAATGATGTAT
AGTTTCCCTGGGCAAACTAAAATTACCCCTCAGGAAATGCCCCCAAAGTGAAATGACAAATGGCATATGAAATGCAAAAT
CCTGCTGGTAAAATCCTTATCCCAAGGCTCATGTAGCTGTGGGACTGGGTTTGATGGCACCGACAGATTTACTAANAGCT
AAAGGCAAATTCATAAAAAAAATNTGGGAAAAATGACAAAATCTGAAAACTGTCTGTTTTAAAAGACAAATTCACAAAAGT
GGAGATAGAGGTTTGTGAATTAGGTGGAAAATCCGACAAATCTATCTCCACTTTTATTTTGTCTCAATATCACCACTTTT
GTGAATTCCCCC
```

Figure S12. Structure (top) and sequence (bottom) of the *X. tropicalis* PiggyBac 451.203.189 DNA transposon which can recruit the p300 co-activator in LETS hybrid embryos. The component annotations from the repeat track are shown.

REFERENCES

1. Wendel JF: The wondrous cycles of polyploidy in plants. Am J Bot 2015, 102:1753-1756.
2. Soltis PS, Soltis DE: Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol 2016, 30:159-165.
3. Ohno S: Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol 1999, 10:517-522.
4. Holland PW, Garcia-Fernandez J: Hox genes and chordate evolution. Dev Biol 1996, 173:382-395.
5. Grant SG: The molecular evolution of the vertebrate behavioural repertoire. Philos Trans R Soc Lond B Biol Sci 2016, 371:20150051.
6. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al: The Atlantic salmon genome provides insights into rediploidization. Nature 2016, 533:200-205.
7. Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC: A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol Phylogenet Evol 2004, 33:197-213.
8. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al: Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 2016, 538:336-343.
9. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol 2014, 31:448-454.
10. Schnable JC, Springer NM, Freeling M: Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A 2011, 108:4069-4074.
11. Song Q, Chen ZJ: Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol 2015, 24:101-109.
12. Bhaumik SR, Smith E, Shilatifard A: Covalent modifications of histones during development and disease pathogenesis. Nat Struct Mol Biol 2007, 14:1008-1016.
13. Perino M, Veenstra GJ: Chromatin Control of Developmental Dynamics and Plasticity. Dev Cell 2016, 38:610-620.
14. Carroll SB: Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 2008, 134:25-36.
15. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al: Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009, 462:315-322.
16. Hontelez S, van Kruijsbergen I, Georgiou G, van Heeringen SJ, Bogdanovic O, Lister R, Veenstra GJ: Embryonic transcription is controlled by maternally defined chromatin state. Nat Commun 2015, 6:10148.
17. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al: Enhancer evolution across 20 mammalian species. Cell 2015, 160:554-566.
18. Matsuda Y, Uno Y, Kondo M, Gilchrist MJ, Zorn AM, Rokhsar DS, Schmid M, Taira M: A New Nomenclature of *Xenopus laevis* Chromosomes Based on the Phylogenetic Relationship to Silurana/*Xenopus tropicalis*. Cytogenet Genome Res 2015, 145:187-191.
19. Weckselblatt B, Rudd MK: Human Structural Variation: Mechanisms of Chromosome Rearrangements. Trends Genet 2015, 31:587-599.
20. Subramanian S, Kumar S: Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res 2003, 13:838-844.
21. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M: Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 2010, 11:R26.

22. Steinberg J, Honti F, Meader S, Webber C: Haploinsufficiency predictions without study bias. Nucleic Acids Res 2015, 43:e101.
23. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al: Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 2013, 5:28.
24. Narbonne P, Simpson DE, Gurdon JB: Deficient induction response in a *Xenopus* nucleocytoplasmic hybrid. PLoS Biol 2011, 9:e1001197.
25. van Kruijsbergen I, Hontelez S, Elurbe DM, van Heeringen SJ, Huynen MA, Veenstra GJ: Heterochromatic histone modifications at transposons in *Xenopus tropicalis* embryos. Dev Biol 2016.
26. Slotkin RK, Martienssen R: Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 2007, 8:272-285.
27. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR: Human mutation rate associated with DNA replication timing. Nat Genet 2009, 41:393-395.
28. Charlesworth B, Morgan MT, Charlesworth D: The effect of deleterious mutations on neutral molecular variation. Genetics 1993, 134:1289-1303.
29. Sasaki M, Lange J, Keeney S: Genome destabilization by homologous recombination in the germ line. Nat Rev Mol Cell Biol 2010, 11:182-195.
30. Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R, et al: Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 2016, 530:171-176.
31. Patel A, Horton JR, Wilson GG, Zhang X, Cheng X: Structural basis for human PRDM9 action at recombination hot spots. Genes Dev 2016, 30:257-265.
32. Nishihara H, Kobayashi N, Kimura-Yoshida C, Yan K, Bormuth O, Ding Q, Nakanishi A, Sasaki T, Hirakawa M, Sumiyama K, et al: Coordinately Co-opted Multiple Transposable Elements Constitute an Enhancer for wnt5a Expression in the Mammalian Secondary Palate. PLoS Genet 2016, 12:e1006380.
33. de Souza FS, Franchini LF, Rubinstein M: Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol 2013, 30:1239-1251.
34. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T: Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res 2014, 24:1963-1976.
35. Villar D, Flicek P, Odom DT: Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet 2014, 15:221-233.
36. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760.
37. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008, 9:R137.
38. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841-842.
39. Georgiou G, van Heeringen SJ: fluff: exploratory analysis and visualization of high-throughput sequencing data. PeerJ 2016, 4:e2209.
40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15-21.
41. Roberts A, Pachter L: Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 2013, 10:71-73.

42. Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Gomez-Skarmeta JL: The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. Methods 2013, 62:207-215.
43. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, et al: Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 2011, 471:68-73.
44. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR: MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc 2015, 10:475-483.
45. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10:R25.
46. Bogdanovic O, Smits AH, de la Calle Mustienes E, Tena JJ, Ford E, Williams R, Senanayake U, Schultz MD, Hontelez S, van Kruijsbergen I, et al: Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet 2016, 48:417-426.
47. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD: A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One 2013, 8:e81148.
48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297-1303.
49. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011, 43:491-498.
50. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al: From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013, 43:11 10 11-33.
51. Slater GS, Birney E: Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 2005, 6:31.
52. Ranwez V, Harispe S, Delsuc F, Douzery EJ: MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 2011, 6:e22594.
53. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T: FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res 2012, 40:W580-584.
54. Charif D, Lobry JR: SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Edited by Bastolla U, Porto M, Roman HE, Vendruscolo M. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007: 207-232
55. Lawrence M, Gentleman R, Carey V: rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 2009, 25:1841-1842.
56. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: Software for computing and annotating genomic ranges. PLoS Comput Biol 2013, 9:e1003118.
57. Mi H, Muruganujan A, Casagrande JT, Thomas PD: Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 2013, 8:1551-1566.
58. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW: CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 2010, 38:D497-501.

CHAPTER FIVE

Dynamics of gene-regulatory networks during embryonic development

Georgios Georgiou
Simon J. van Heeringen
Gert Jan C. Veenstra

ABSTRACT

Embryonic development is regulated by dynamic patterns of gene expression, which are orchestrated through the action of complex gene regulatory networks. Recent genome-wide studies have revealed that these gene regulatory networks are highly complex. Transcription factor (TF) proteins interact extensively with cis-regulatory elements and influence gene expression. Here we aim to gain insights to the gene regulation by combining diverse genome-wide data. To reveal and describe the regulatory program of individual genes, we have developed an ensemble learning method to construct gene regulatory networks. Our method integrates binding of the p300 (*Ep300*) co-activator, transcription factor expression and transcription factor motif scores to infer gene-regulatory interactions. Using a combination of benchmarks, such as transcription factor ChIP-seq data and experimentally validated interactions, we show that the ensemble approach outperforms individual methods for predicting regulatory interactions. We applied our method in X.*tropicalis* embryos, spanning developmental stages from blastula to tailbud embryos, with the aim to study the gene regulatory network dynamics during vertebrate development. Using the network information to predict influential TFs, we found transcription factors driving developmental transitions and we identified sub-networks associated with important biological processes. Finally, we inferred spatially resolved transcription factor networks in gastrula-stage embryos. Our work shows that different genome-wide assays of regulatory potential can be complementary to each other and that combining information can improve network inference.

BACKGROUND

Gene regulation is the process that controls the development of a multicellular organism from a single fertilized egg. It is coordinated by an interplay between chromatin and transcription factors (TFs). TFs work together as gene regulatory networks (GRNs). Complex GRNs, encoded in the genome, regulate the precise spatial and temporal control of gene expression. TFs bind to regulatory regions, such as enhancers and promoters (Heintzman et al. 2009; Stender et al. 2010; Heinz et al. 2010). They bind by recognizing small DNA sequences, also known as motifs and influence the expression of their target genes (Spitz and Furlong 2012). They can function as activators or repressors and recruit other co-activators or co-repressors and RNA polymerase II (Matsui et al. 1980; Zawel and Reinberg 1993). The coactivator p300 is encoded by the gene *ep300* and is recruited to enhancers by TFs (Eckner et al. 1994; Chan and La Thangue 2001). p300 facilitates the deposition of MLL4-dependent H3K4me1, which subsequently boosts the deposition of p300-mediated H3K27ac (Wang et al. 2017). Therefore, the presence of p300 accurately pinpoints enhancer location and can be used as an indicator of enhancer activity (Blow et al. 2010; Visel et al. 2009).

GRNs play a significant role in development and pluripotency by controlling which genes are activated or repressed. Therefore, having an accurate GRN is an essential step towards understanding gene regulation. Small-scale GRNs can be constructed by experimental approaches. However, because these approaches are focused on individual genes or TFs, they tend to be expensive and time-consuming (Koide, Hayata, and Cho 2005; Charney et al. 2017; Loose and Patient 2004). Nowadays, with computational approaches and advances in high-throughput sequencing, inferring interactions between TF and a gene and therefore a GRN, can be achieved with relatively low cost and time (Liu 2015; Lee and Tzou 2009; Delgado and Gómez-Vela 2018).

GRNs can be based on gene expression measurements or on other regulatory data. Expression-based networks measure a degree of similarity in expression pattern between pairs of genes in different conditions, such as across time-series, tissues and knockouts. They tend to be undirected and only assume that there is a functional relationship between the pair (Serin et al. 2016). By combining gene expression with other information, such as transcription factor binding, long-range interactions, and chromatin accessibility, TFs can be linked to their target genes with directionality. GRNs using a wide range of information have been inferred for many organisms, such as *E. coli*, *S. cerevisiae*, and *Drosophila* (Ernst et al. 2008; Harbison et al. 2004; Marbach, Roy, et al. 2012).

Elucidating GRNs can have important implications for studies of embryonic development. They can be used to identify putative functions for uncharacterized genes and narrow down the potential interactions between TFs and genes (Marbach, Roy, et al. 2012). One of the model organisms used to study embryonic development is the amphibian *Xenopus tropicalis*. This organism shares several anatomical, physiological and genetic similarities with humans (Wheeler and Brändli 2009; Schmitt, Gull, and Brändli 2014; Hempel and Kühl 2016). These similarities can help us understand vertebrate genes, study functional regulation, contribute to drug discovery and explore the underlying molecular mechanisms of congenital diseases (Duncan and Khokha 2016; Hempel and Kühl 2016; Blum and Ott 2018). Finally, we have available genome-wide maps of various histone modifications, binding of the *Ep300* co-activator and a high-resolution expression profiling of *Xenopus* during development (Hontelez et al. 2015; Owens et al. 2016). These data, together with a non-redundant database of TF motifs (Weirauch et al. 2014) make an extensive collection of information that can be used to infer gene regulatory interactions.

Here we present a network-based approach to help us study gene regulation development by inferring novel interactions and getting new insights in important biological processes. Using genome-wide functional data, we constructed regulatory networks during early embryonic development in *Xenopus tropicalis*. We combined binding of the p300 (*Ep300*) co-activator, transcription factor expression and transcription factor motifs score in an ensemble approach to infer regulatory binding networks. The edges in such a network represent binding of a corresponding TF. We assess the three initial networks and ensemble network using ChIP-seq data of transcription factor binding and literature-based validation. Using TF ChIP-seq as a benchmark, we saw that *Ep300* ChIP-seq signal has the most discriminatory power out of the three data types. However, the ensemble approach improved upon the individual predictions. Using literature-based interactions as validation data, the ensemble method predicted the majority of the experimentally validated interactions (AUC: 0. 92 and 0.98).

Our inferred networks predict the binding of TFs to enhancers in gene loci. Looking at the binding of Eomes in the *irx1* gene locus, our approach was able to predict the majority of the binding sites (Figure 2D). However, binding does not always imply regulation. Under the assumption that expression levels of TFs and their targets tend to correlate, we incorporate expression data to find functionally related genes and infer "regulatory networks". To determine the biological relevance of the inferred networks we use gene ontology (GO) and as *Xenopus* Anatomy Ontology (XAO) annotations (Ashburner et al. 2000; The Gene Ontology Consortium 2017; Segerdell et al. 2013). Using network clustering methods, we identified eight communities with nodes densely connected with each other and sparsely connected with the rest. We calculated the enrichment of those communities in GO and XAO annotations and found a

community enriched with genes associated with pattern specification and regionalization processes. Using the network structure and expression time-course data we scored each transcription factor based on their influence on their target genes expression. For the developmental transition from the blastula to the gastrula stage, the top predicted influential TFs were known to be important for gastrulation. Using the network structure and expression time-course data we scored each transcription factor based on their influence on target gene expression. We combined our early gastrula (stage 10.5) network with regionalized mRNA expression profiles from early gastrula embryos to determine more spatially resolved TF-TF regulatory networks. In the animal cap network, we identified TFs known to be important in the specification of neuronal versus non-neuronal ectoderm. In the ventral, lateral and dorsal marginal zone networks, we found that T-box transcription factors (Tbxt, Vegt, Eomes) have an essential role. The T-box transcription factors are known to be important in mesoderm formation and differentiation. We found the Foxa4 transcription factor present in the marginal zone and vegetal TF networks, but not in the animal cap network. Foxa4 is known to cooperate with T-box transcription factors in dorsal mesoderm formation. Finally, in the vegetal network, we found Vegt and Sox17a to be among the most important transcription factors.

RESULTS

Inference and validation of stage-specific binding networks

We have used chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) of the histone acetyltransferase *Ep300* to determine putative cis-regulatory elements. We combined *Ep300* ChIP-seq peaks for five developmental stages in X.*tropicalis* (Nieuwkoop-Faber stage 9, 10.5, 12.5, 16 and 30) (Nieuwkoop and Faber 1994),spanning blastula to tailbud embryos (Hontelez et al. 2015). This resulted in a total of 126,578 cis-regulatory regions. We linked these cis-regulatory regions to genes using the regulatory domains as defined by the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al. 2010).

Based on the premise that expressed TFs will bind their cognate motifs inside active regulatory elements, we inferred gene-regulatory interactions. Using an ensemble approach and mean rank aggregation, we combined enhancer activity, TF motif scores, and TF expression into a single model (Figure 1A). As a measure for enhancer activity, we used the normalized *Ep300* ChIP-seq peak intensity (reads per kilobase per million reads; RPKM) (See Methods). This resulted in an enhancer-based network. This network has edges from enhancers to genes and the weights represent the normalized ChIP-seq signal. To determine the TF motif scores, we scanned the putative enhancers for 480 motifs associated with 942 X.*tropicalis* transcription factors (See Methods). The TFs motif score edge weights represent the maximum score for each motif per enhancer region. To quantify the expression of the TFs we used RNA

sequencing data (RNA-seq) from the corresponding developmental stages of interest (See Methods). The TFs expression network has edges from TFs to genes and the weight represents the expression of each TF in the corresponding stage. We represented the datasets as three individual, directed and weighted networks (Figure 1A). To create an aggregated regulatory enhancer network, we assigned a score for every TF in an enhancer. The ensemble score was calculated by combining the three individual edge weights using mean rank aggregation. However, a gene can have multiple enhancers in its locus, which will result in multiple weights for each TF-gene combination. For a gene with multiple enhancers, we used the highest score among all its enhancers. This resulted in five stage-specific directed consensus networks with 15,445,032 (942 TFs * 16,396 genes) TF-gene weighted edges each.

To assess the quality of the TF-enhancer networks we used experimental ChIP-seq data to evaluate how well these networks could predict TF binding. We collected public *X. tropicalis* stage 10.5 ChIP-seq data of eight transcription factors: Eomes (Gentsch et al. 2013), Tbxt (Brachury) (Gentsch et al. 2013), Vegt (Gentsch et al. 2013), Foxh1 (Chiu et al. 2014), Gsc (Yasuoka et al. 2014), Otx2 (Yasuoka et al. 2014), Smad2/3 (Yoon et al. 2011) and Sox2 (unpublished). We also included beta-catenin (Nakamura et al. 2016), the downstream effector of the Wnt-signaling pathway (MacDonald, Tamai, and He 2009). We mapped the reads to the genome and identified peaks that overlapped with our collection of *Ep300*-based regulatory regions. Using this experimental evidence of binding as a reference, we evaluated the three individual networks and the aggregated consensus network for stage 10.5 (Figure 1B). Out of the three methods to determine edge weights we tested, *Ep300* ChIP-seq signal has the most discriminatory power (ROC AUC: 0.855), while the other two showed lower performance (motif score ROC AUC: 0.657, TF expression ROC AUC: 0.658). However, the ensemble approach outperforms all individual methods for predicting transcription factor binding (ROC AUC: 0.862).

To assess the quality of the TF-gene stage 10.5 regulatory network, we used two different benchmarks for regulatory interactions. First, we used gene co-citation data. The co-citation dataset consists of *Xenopus* genes and TFs co-cited in the same paper, based upon data from Xenbase (Karimi et al. 2018). The assumption here is that TFs and their target genes will likely be cited together. However, the gene co-citation dataset will capture many other types of relationships; hence, this set will not contain only true positive interactions within this specific context. Second, we used a collection of experimentally validated interactions from curated *Xenopus* mesoderm networks (Koide, Hayata, and Cho 2005; Charney et al. 2017). These mesendoderm networks were assembled from large-scale literature curation and contained all known TF-gene regulatory interactions known to be important in mesendoderm and early endoderm in X.*tropicalis* and X.*laevis*. Using these two datasets, we show that our ensemble method had good performance in predicting interactions using the co-citation

"gold standard" (ROC AUC: 0.74) (Supplementary Figure 1) and excellent performance in predicting experimentally validated interactions (ROC AUC: 0.92 and 0.98) (Figure 1C). Overall, the evaluations demonstrate the predictive power of our method to reconstruct TF-gene networks and confirm their biological significance.

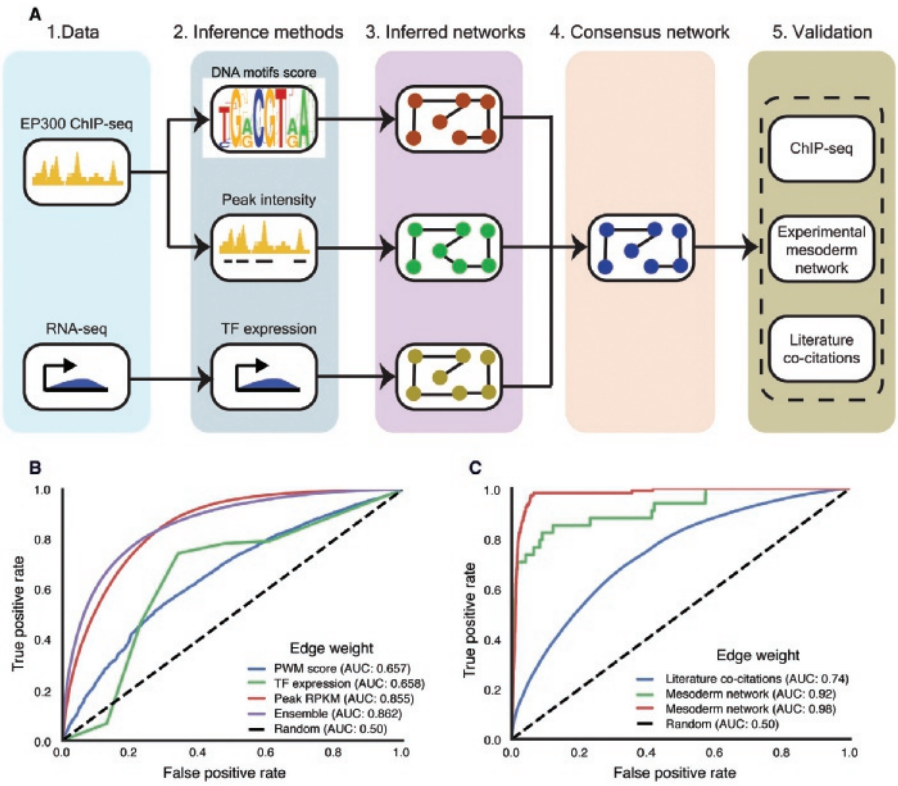


Figure 1. Overview of the method for generation and evaluation of the binding network (A) Generation of the binding network using our Ensemble method involved the following step (from left to the right). **(A.1.)** We used *EP300* ChIP-seq and RNA-seq data from 5 developmental stages as input for each network. **(A.2.)** *Ep300* ChIP-seq was used to identify the cis-regulatory regions. Regions were scanned for motifs using our vertebrate motif database. We also obtained the *EP300* ChIP-seq signal intensity of each region as a measure for enhancer activity. RNA-seq data were used to find which transcripts are (higher) expressed in each stage. **(A.3.)** We ended up with an edge weight for each measurement. The motif score measures how well a motif matches the underlying sequence. The *Ep300* peak RPKM is a measure for enhancer activity. The TF expression (TPM) as measured by RNA-seq represents the activity of each TF. **(A.4.)** To construct the binding network, we calculated the mean ranked edge weight across the three networks. **(A.5.)** The consensus network was validated using experimental and literature data **(B)** We validated the stage 10.5 binding network using TF ChIP-seq data. Out of the three edge weights, *Ep300* peak RPKM (blue) is the most discriminate with an AUC of 0.86. The PWM score (green) and TF expression (red) follow with performance better than random and AUC of 0.67 and 0.66 respectively.

The ensemble approach (purple) outperforms all three individual edge weights with AUC of 0.87 **(C)** Literature validation for the stage 10.5 binding network. We used literature co-citation data and experimentally validated mesoderm networks as validation for the ensemble method and how well it predicts binding of TF in a gene locus. Our approach had good performance at the co-citations data (blue) with AUC of 0.77. Using the two mesoderm networks as validation, the approach has excellent performance in predicting known interactions with AUC of 0.90 (green) and 0.93 (purple).

Structural characteristics of the early gastrula stage embryos binding network

Having evaluated the method, we sought to examine the structural characteristics of the predicted networks. Using our ensemble approach, we predicted TF binding sites in *X. tropicalis* embryos at NF stage 10.5. We selected the top 101,756 predicted edges (estimated FDR of 30%). The number of targets per transcription factor varies from 3 to 3,643, with an average of 988 targets (median: 742) (Figure 2A). The TFs with the highest number of predicted targets at this stage are Foxa4 (3,643), Vegt (3,538), Sox3 (3,422), Sox11 (2,949), Tbxt (2,976) and Sox2 (2,939). All these TFs are important developmental regulators in early *X. tropicalis* embryogenesis. Foxa4 is the most abundantly expressed foxa gene at this stage in development (Charney et al. 2017). It is required for correct specification of the notochord and is involved in anterior-posterior patterning of the neural plate (Murgan et al. 2014). Vegt is a maternal TF required for mesoderm and endoderm formation in the embryo (Xanthos et al. 2001; Clements, Friday, and Woodland 1999; Howard et al. 2007). It interacts with beta-catenin to induce mesoderm formation (Kofron et al. 1999; J. Zhang et al. 1998; Fukuda et al. 2010). The SoxB1 genes Sox2 and Sox3 are highly expressed in the pluripotent blastula stage in later stages are localized to the neural ectoderm (Buitrago-Delgado et al. 2018).

The in-degree of target genes varies between 1 and 87 with an average of 20 incoming edges (median: 15) (Figure 2B). The genes that have the highest number of predicted regulators include *irx1*(82), znf608 (83), znf703 (80), and sp5l (79). These genes generally have hundreds of enhancers in their vicinity, with the exception of sp5l, which has 50 enhancers in its locus. Due to the implementation of our network, genes with many of enhancers have a higher probability of being regulated by many TFs. The distribution of in- and out-degrees of the predicted network resembles a scale-free distribution, as was previously reported for other biological networks (Marbach, Roy, et al. 2012).

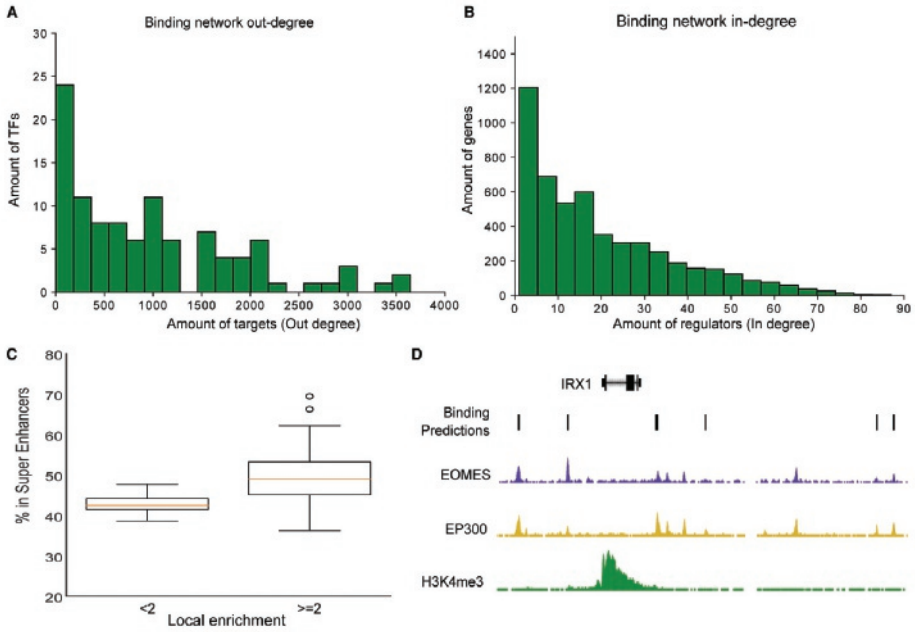


Figure 2. Characterization of the stage late gastrula binding network (A) The out-degree distribution of TFs in the binding network follows the power-law distribution. Shown is the distribution of out-degree (number of target genes a TF can regulate). **(B)** The in-degree distribution of target genes in the binding network follows a power-law distribution. Shown is the distribution of in-degree (number of regulators a gene can have). **(C)** Local enrichment of TF binding. Shown is the percentage of TFs bound in super-enhancers. TFs which are found to be enriched (n=87; enrichment = fraction of local enhancers bound by TF / fraction of random enhancers bound by TF) in loci with more than five enhancers are often bound in SE regions. On the other hand, TFs with binding that is not locally enriched around genes genome-wide (n=24) are less often found in SE. **D)** Genome Browser screenshot from the *IRX1* gene locus with late gastrula (stage 10.5) ChIP-seq enrichment of H3K4m3 (green), *Ep300* (yellow), and Eomes (purple). Our method predicted (black bars) the majority of true EOMES binding sites as identified by ChIP-seq (purple track).

TFs tend to bind to multiple enhancers around their target genes (Long, Prescott, and Wysocka 2016). Therefore, genes regulated by a specific TF will be enriched for TF binding at multiple enhancers in their locus relative to random enhancers. Enhancers also tend to cluster around key developmental regulator genes in super-enhancers (SE) (Parker et al. 2013; Whyte et al. 2013). We wondered if our predicted network would capture the enrichment of multiple binding events of TFs in SEs. We therefore used the binding network to look for TFs locally enriched in regulatory regions around genes. For the analysis we used only genes associated with a minimum five enhancers and TFs predicted to be bound in at least 10 enhancers. Out of the 104 TFs in the network, 87 were found to be locally enriched relative to randomly selected enhancers (adj. pval < 1e-5). We examined the binding location of these TFs and we found that

49.42% is bound in SE regions. This indicates that being in an SE plays a significant role in the local enrichment. In comparison, non-enriched TFs bind in SE regions less frequently (42.76%, Table 1). For example, Eomes is predicted to bind at six locations in the *irx1* locus SE. Genome-wide, multiple binding events of Eomes around loci is found to be 1.76 times enriched and a 45.92% of its binding was within SE regions. It appears that our binding network recapitulates the known clustering of TF binding events in super enhancers.

In summary, we created a binding network by combining three different measures of gene regulation at enhancers using mean rank aggregation. Our network exhibits similar structural characteristics as other biological networks. The TFs with the most predicted targets are known to be important regulators at this developmental stage. Finally, looking at SEs and local enrichment, we saw that genes associated with SEs generally have a higher local enrichment of bound TFs.

Construction and structural characteristics of the early gastrula stage regulatory network

Our inferred networks predict the binding of a TF to an enhancer in a gene locus. However, binding does not always imply regulation. To overcome this limitation, we incorporate co-expression network information. Using the rationale that co-expression can imply regulation, we construct a GRN by integrating high-resolution developmental gene expression data with our binding network (Collart et al. 2014; Owens et al. 2016) (Figure 3A). For all TF-target gene combinations, we calculated the pairwise correlation between their expression levels from 4.5 hpf to 66 hours post fertilization (hpf) (Figure. 3A). Subsequently, we filtered the binding networks by keeping only edges from gene-pairs with high gene expression correlation (Pearson r >=0.70). The resulting network contains 4,399 edges and is referred to as the "regulatory network".

Similar to the binding network, the TF out-degrees of the regulatory network exhibit an almost scale-free distribution. The number of targets per transcription factor varies from 1 to 250, with an average of 46 targets (median: 27) (Figure 3B). At the top, with the most targets, we have Ybx1 (250), Tef (226), Id2 (224) and Hif1a (219). Following the same pattern, the in-degree of the targets follows an almost scale-free distribution, with the regulators per gene to be between 1 to 21, with an average of three targets (median: 2) (Figure 3C). The genes regulated by many TFs include zswim4 (21), *ventx*2.1 (17), vegt (17), *ventx*1.1 (18) and ets2 (19).

Biological networks usually have a modular structure (Shen-Orr et al. 2002; Milo et al. 2002; Jeong et al. 2000; Spirin and Mirny 2003). They are organized in compartments called communities (Blondel et al. 2008; Fortunato 2010; Girvan and Newman 2002; Clauset, Newman,

and Moore 2004; Rosvall and Bergstrom 2008; Lancichinetti and Fortunato 2009; Lancichinetti et al. 2011). Communities represent nodes densely connected with each other and sparsely connected with the rest of the network. Communities can have their own role and function in the network. They can correspond to sets of genes associated with the functions or processes, therefore, detecting communities is an important step towards understanding the structure and organization of the network. To identify communities, we used edge betweenness as the centrality measure and the fast greedy modularity optimization algorithm for finding community structure (Freeman 1978; Brandes 2001; Clauset, Newman, and Moore 2004). Genes with high betweenness are likely to act as hubs between communities in the network. We examined the community structure of our networks and identified eight distinct communities with genes highly interconnected between them and with lower connectivity with other clusters (Supplementary Figure 2, Figure 3E). Using *Xenopus* Anatomy Ontology (XAO) and Gene Ontology (GO) annotations, we calculated the enrichment of those communities in functional (Table 2) and anatomical terms (Table 3) (Ashburner et al. 2000; The Gene Ontology Consortium 2017; Segerdell et al. 2013). We identified a community (Community 1) that contained the TFs Otx1, Otx2, Foxa4, Eomes, Tbxt, Tcf7l1, Foxh1.2, Sox3 and Gsc, which was enriched in pattern specification and regionalization processes (Supplementary Figure 2). The mesodermal markers, Eomes and Tbxt, pinpoint this community as potentially being involved in the Spemann-Mangold organizer, a cluster of cells which induce the dorsal-ventral axis and neural tissues (Crease, Dyson, and Gurdon 1998). Based on a visual assessment, we saw that the majority of the large communities (>200 nodes) had a highly interconnected structure. The only exception was "Community 1". That lead us to believe that it could contain more fine-grained sub-structures. We re-clustered "Community 1" and we identified sub-communities within it (Figure 3F). These sub-communities were also enriched in anatomical (Table 4) and functional (Table 5) terms.

We assessed the biological relevance of the binding and regulatory networks based on the tendency of genes targeted by similar TFs (co-regulated targets) to exhibit similar functional properties or being localized to similar tissues. To evaluate and compare the performance of the networks, we calculated the enrichment of co-regulated genes in GO and XAO annotations for each network. We considered two genes as co-regulated if they share more than 50% of the regulators (Jaccard index >= 0.50). We repeated the same process on randomized networks (n=1000) while maintaining the same network structure as the original networks. Then we compared the average enrichment in GO and XAO terms between our networks and randomized networks. As expected, our networks were enriched in both annotations compared to the random networks. We noticed that the enrichment is higher in the early stages, likely due to less specific regulatory predictions in later stages due to the cellular complexity of the embryo. Surprisingly, we saw that co-expression information did not result in stronger

functional enrichment. The binding network had higher enrichment in all stages in XAO and GO terms, with the exception of the stage 10.5 regulatory network in GO terms.

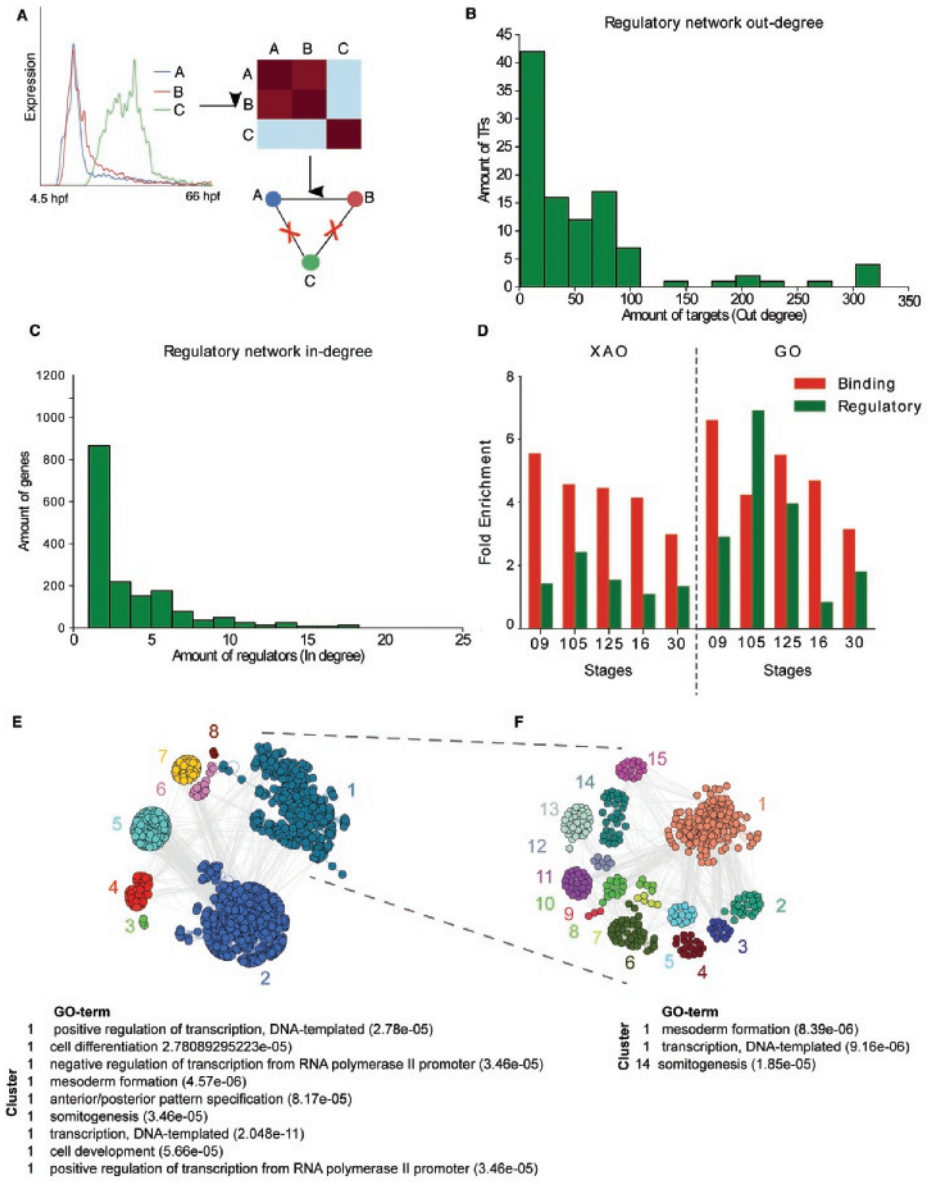


Figure 3. We incorporate expression data to find functionally related genes and infer regulation. (A) Overview of the hypothesis we used to infer GRNs. Under the assumption that the mRNA level of TFs and its targets tend to correlate, we perform genome-wide clustering of gene expression profiles from 4.5 hpf to 66 hpf.

Expression correlation was combined with the binding network to confirm gene regulation. Edges were kept only if the corresponding nodes correlated (>=0.7). (A: FOXI4.2, B: T, C: CRISP1). **(B)** Out-degree of TFs in the stage 10.5 regulation network follows power-law distribution. Shown is the distribution of out-degree (number of target genes a TF can regulate). **(C)** In-degree of target genes in the stage 10.5 regulation network follows the power-law distribution. Shown is the distribution of in-degree (number of regulators a gene can have). **(D)** Enrichment of binding and regulatory networks for each developmental stage in *Xenopus* Anatomy Ontology (XAO) (left) and gene ontology (GO) (right) annotations, relative to randomized networks. All networks show high enrichment in both data sets. Overall, binding networks were higher enriched compared to the regulatory networks. **(E)** The early gastrula stage regulatory network is visualized using community structure. Communities were identified using the edge betweenness as the centrality measure and the fast greedy modularity optimization algorithm for finding community structure. Shown is the functional enrichment of Community 1. **(F)** The "Community 1" is visualized with the identified sub-communities. Sub-communities were identified using the edge betweenness as the centrality measure and the fast greedy modularity optimization algorithm for finding community structure. Shown is the functional enrichment of "Community 1" and "Community 14".

In summary, we combined the information from the binding networks with gene expression data to create a regulatory network. Our network exhibits similar structural characteristics as our binding networks and as other biological networks. Using the edge betweenness as the centrality measure and the fast greedy modularity optimization algorithm we identified eight communities. Out of the eight communities, "Community 1" was the most interesting because it was found to be enriched for genes related to pattern specification and regionalization processes. Finally, we assessed the biological relevance of our networks based on the tendency that co-regulated target genes exhibit similar functional properties or being localized to similar tissues. We saw all networks were enriched in functional terms compared to random networks, but binding networks had a higher enrichment compared to the regulatory networks. Due to the complexity of the embryo, the enrichment was higher in the networks of early-stage embryos.

Predicting key TFs involved in developmental transitions

Previously it has been shown that the information encoded in GRNs can be used to identify cell fate regulators (Rackham et al. 2016; Morris et al. 2014). We hypothesized that a similar method could be used to identify TFs that drive developmental transitions. To identify TFs responsible for gene expression changes we calculated the "influence score" using an adaptation of the Mogrify method (Rackham et al. 2016). TFs get a score based on the expression change of differentially expressed targets in their local network (up to the third degree (See Methods). In this approach TFs that regulate a few genes that are highly differentially expressed are expected to have a higher score compared to genes with many predicted targets that are mostly not differentially expressed (Figure 4A).

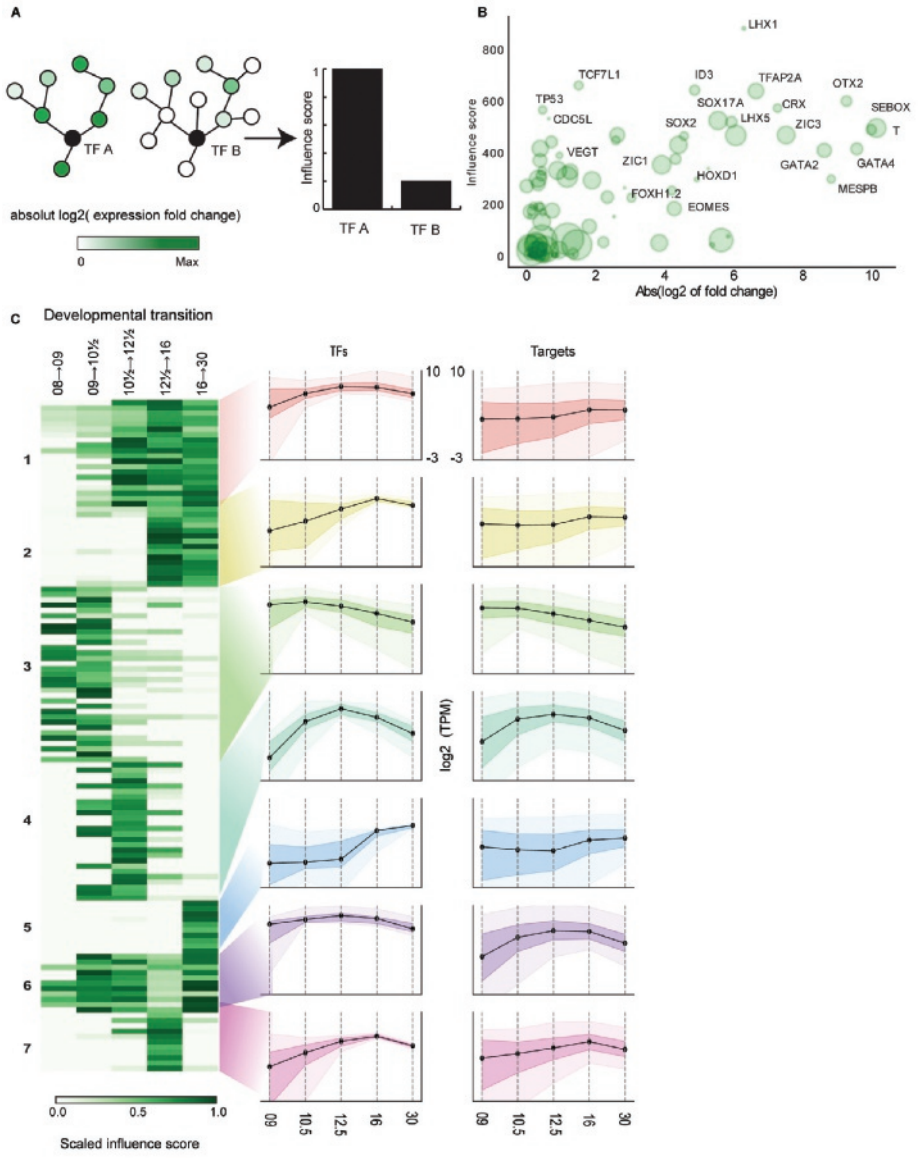


Figure 4. Transcription factors that drive developmental transitions. (A) The outline of the method used. For each transcription factor, we constructed a local network up to the third degree. For every target gene in the local network, the expression change is calculated. Transcription factors regulating more differentially expressed genes will have a higher influence score compared to those who regulate fewer or barely differential. **(B)** There is not a uniform linear relationship between the expression change and the influence score. Expression change and influence score values are for the transition from stage 9 to stage 10.5. The size of the point corresponds to the direct targets of the respective TF in stage 10.5. **(C)** Stage-specific influential transcription factors for early as well as for late development and through

all the five developmental stages. Shown (left panel) is the influence score of each transcription factor. We identified seven clusters of transcription factors with similar influence score patterns. Transcription factors peak in expression (center panel) at the stage with their highest influence score. Their target genes (right panel) follow similar gene expression patterns.

Early in development, the embryo transitions from the blastula to the gastrula stage, where it establishes the anteroposterior and dorsoventral axes and the three primary germ layers are formed and organized. It is known that TFs such as Vegt, Tbxt, and Otx2 are essential for germ layer formation. We calculated the influence score for all TFs in the developmental transition from NF stage 9 to stage 10.5. In Figure 4B all TFs are plotted with the predicted influence score as a function of the expression change (stage 10.5 compared to 9). Tcf7l1, Otx2, Sox17a, Tbxt, Sox2, and Vegt are amongst the top predicted influential TFs (Figure 4B). These TFs are known as key TFs for gastrulation. We observed that there is not a uniform linear relationship between the expression change and the influence score. Some transcription factors, such as Vegt, Tp53, Tcf7l1, and Cdc5l, have high influence score in the transition from stage 9 to stage 10.5, but their expression barely changes. This effect can be explained by the amount of highly differentially expressed direct and indirect targets. It indicates that the expression change by itself would not be sufficient to predict these TFs as important regulators for this transition.

To study the dynamics of gene regulation and how the influence of TFs changes during development we calculated the influence score in five developmental transitions spanning the stages from the blastula to tailbud embryo (NF stage 8 to 9, 9 to 10.5, 10.5 to 12.5, 12.5 to 16 and 16 to 30). As the networks have different sizes and as the range of influence scores can vary between them, we scaled the scores from zero to one to enable comparison. To identify TFs following the same pattern of influence we performed clustering on influence scores using k-means clustering with the Euclidean distance. We identified seven clusters, exhibiting distinct patterns in influence score during the developmental transitions (Figure 4C). We identified clusters with stage-specific influential transcription factors for early as well as for late development and through all the five developmental stages. As expected, transcription factors and their targets follow a similar pattern in expression because of the properties of the network. Interestingly, the influence score of TFs follows the same trend as their expression, with the TFs peak in expression at the stage with their highest influence score (Figure 4C).

Transcription factors regional network in the early gastrula stage embryos

Our predicted networks captures regulatory interactions based on whole-embryo data. To determine more spatially resolved regulatory networks we combined our stage 10.5 network with regionalized mRNA expression profiles from early gastrula embryos (Blitz et al. 2017). We focused on TF-TF interactions and filtered network edges based on expression (TPM >= 40) in the following regions: animal cap (AC), vegetal mass (VM), lateral marginal (LM), dorsal

marginal (DM) and ventral marginal (VG) zones (Fig. 5). For the animal cap, we identified 11 TFs that were expressed and represented in our network. A dense subnetwork is formed by Klf17, Foxi4.2, Tfap2a, Tfap2c and Lhx5. The transcription factors Tfap2a and Tfap2c play an important, evolutionary conserved role in the ectodermal lineage. Tfap2a target genes in *Xenopus* include epidermal as well as neural crest genes (Luo et al. 2002, 2003). In zebrafish, these AP2alpha factors are required ectoderm derivatives such as neural crest (W. Li and Cornell 2007). Foxi4.2 is required for correct ventral specification of the early head ectoderm (Matsuo-Takasaki, Matsumura, and Sasai 2005). Not much is known about the role of the Kruppel-like factor Klf17 in early *Xenopus* development. It is both maternally and zygotically expressed and is enriched at the animal pole in early embryos (Gao et al. 2015). Later in development it is expressed in the cement gland, hatching gland and ventral blood islands. Lhx5 expressed in the entire ectoderm and is likely involved in the development of the nervous system (Peng and Westerfield 2006; Toyama et al. 1995). One possible hypothesis based on the known functions of the genes in this core animal cap network is that these factors are important in specification of neuronal versus non-neuronal ectoderm.

In the ventral, lateral and dorsal marginal zone networks, the T-box transcription factors, Vegt, Eomes, and Tbxt, have an essential role. They are three of the highest expressed transcription factors of the stage 10.5 network and ones with the most targets. This finding highlights their importance in mesoderm formation and differentiation, which corresponds with their well-described role (Kofron et al. 1999; J. Zhang et al. 1998; Ryan et al. 1996; Fukuda et al. 2010; Knezevic, De Santo, and Mackem 1997; Conlon et al. 1996). As expected, organizer TFs such as Gsc, Otx2, Lhx1 and Crx are found in the dorsal but not ventral or lateral marginal zone networks. One other crucial transcription factor that is present in all marginal zone networks is Foxa4. Foxa4 is zygotically expressed, with an increasing expression level during gastrulation. It is present in all three marginal zone networks, however, only in the dorsal marginal zone the number of its targets increases significantly. Foxa4 is known to cooperate with T-box transcription factors in dorsal mesoderm formation and is confirmed by our network (Murgan et al. 2014).

The most important transcription factors in ventral marginal zone network are Vegt, Foxa4, Sox17a. Our findings go along with literature where Vegt and Sox17a are known to be required for embryonic endoderm development (Howard et al. 2007; Engleka, Craig, and Kessler 2001). In addition, several Gata TFs, which are important regulators of endoderm formation are also present in the ventral marginal zone network (Charney et al. 2017).

DISCUSSION

We reconstructed a TF binding network for X.*tropicalis* on basis of genome-wide p300 binding ChIP-seq and RNA-seq data using an ensemble approach. We examined the performance of different approaches (p300 ChIP-seq signal, transcription factor expression, and motif scores) in inferring gene regulatory interactions. We found that, out of these three metrics, p300 ChIP-seq signal had the highest discriminative power leading to the hypothesis that transcription factors located under a robust p300 ChIP-seq peak are expected to be functional. Transcription factor expression and motif score had lower discriminative power. The importance of integrating multiple data types into a single model for predicting gene regulatory networks has been already proven to be significant (Marbach, Costello, et al. 2012). Likewise, we observed that data are complementary to each other with their aggregation improving the prediction. The ensemble approach limits the bias from the different datasets and provides improved discriminative power.

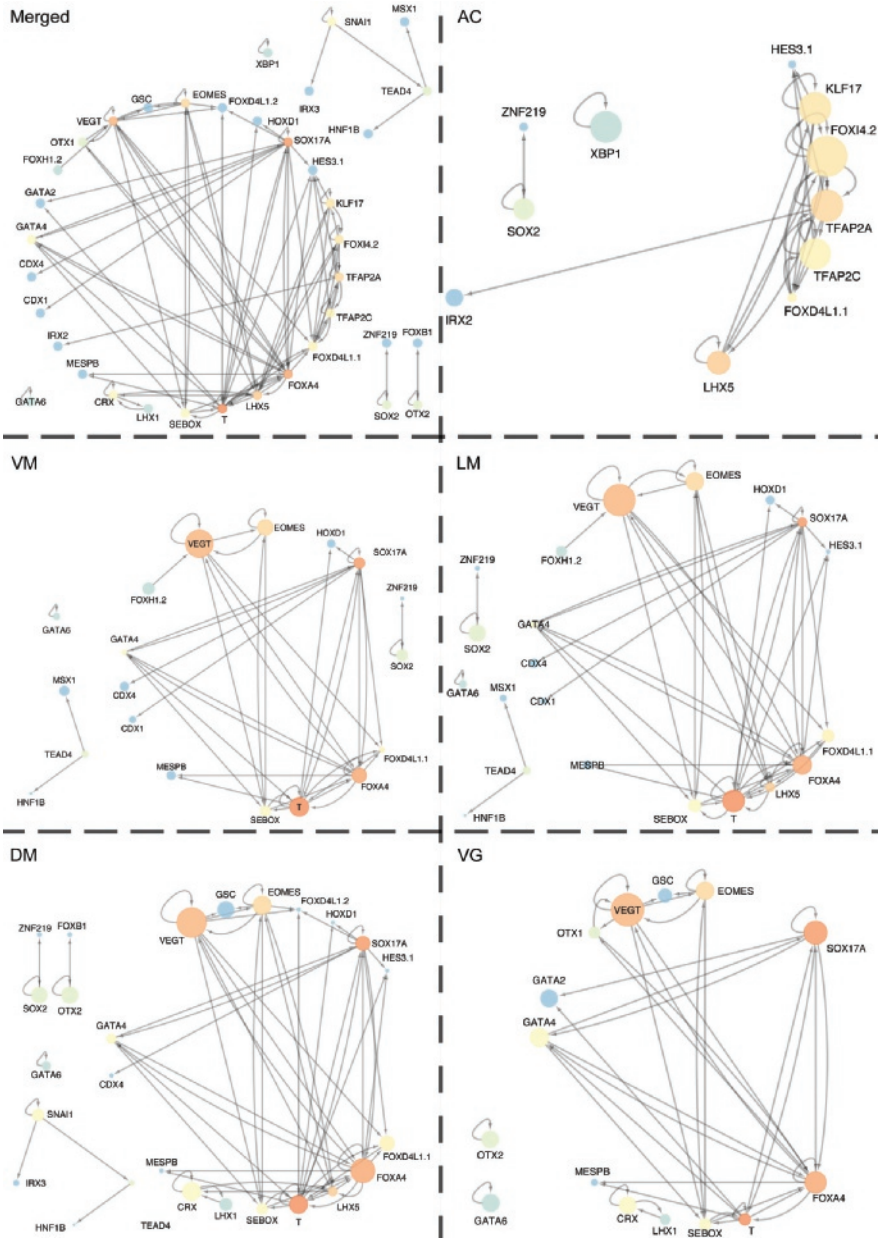


Figure 5. Transcription factors regional networks in the early gastrula stage embryos. Shown are the merged TF-TF network and regional transcription factor networks for the animal cap (AC), vegetal mass (VM), lateral marginal (LM), dorsal marginal (DM) and ventral marginal (VG) zones. In the regional networks, the color of the node corresponds to the number of edges (in- and out-going) on a color scale from blue to red. The blue color represents a transcription factor having fewer edges and red a transcription factor having more. The size of the node indicated the expression level of the transcription

factor in the corresponding region. For each regional network, we identified transcription factors known to be important for that region.

Our network predicts binding of 942 TFs with a known binding motif to the cis-regulatory elements of 16,396 genes. At an estimated 30% FDR we predict 173,000 to 287,000 edges per stage. Looking at the network properties, we saw that the in- and out-degree of the inferred networks resembled a scale-free distribution. This was previously reported also in other biological networks (Marbach, Roy, et al. 2012; Albert 2005). While the predicted binding networks implicitly model regulation to some extent, the inferred edges may be non-functional, in the sense that they do not represent true regulatory relationships (X.-Y. Li et al. 2008). Under the assumption that mRNA levels of TFs and their targets tend to correlate during development, we used RNA-seq data to find co-expressed TFs and genes. Using expression data is a common approach to infer regulatory interactions. Several methods base their predictions on the similarity of expression patterns between transcription factors and genes (Obayashi and Kinoshita 2009; Prill et al. 2010; Butte and Kohane 2000; Margolin et al. 2006). Here we speculated that having an edge from a TF to a gene and good correlation in expression means that the TF is regulating that gene. Based on this premise we constructed regulatory networks for five distinct developmental stages. We saw that our networks had subnetwork structures, which we refer to as communities. These correspond to densely connected nodes, which we found to be associated with anatomical and functional terms. These results go along with what is reported in other developmental and biological GRNs (Oliveri and Davidson 2007; Davidson and Levine 2008; Alcalá-Corona et al. 2016; Marbach, Costello, et al. 2012). We confirmed the relevance of the networks by comparing the enrichment of co-regulated genes with randomized networks. Binding networks had notably higher enrichment compared to the regulatory networks. Using the GRNs in combination with expression data, we scored transcription factors based on their importance in developmental transitions. Similar methods were previously used to identify cell fate regulators (Rackham et al. 2016; Morris et al. 2014). For each developmental transition, we identified key TF hypothesized to drive the transitions. Clustering the TFs based on their influence score resulted in seven TFs clusters. These clusters were dynamic through development. Finally, using our networks and spatial expression data, we constructed spatial transcription factor networks for the animal cap, vegetal mass and lateral, dorsal and ventral marginal zones. We identified transcription factors essential for gene regulation in these zones and our findings correspond with literature.

Limitations and future directions

Although our method provided valuable results and the foundation to infer and study gene regulatory interactions, there are also limitations to this genome-wide approach. We linked cis-regulatory regions to genes by assigning a gene locus based on GREAT regions. This method has its limitations as it associates only proximal regions to genes and assumes that longer genes will correspond to more cis-regulatory regions. Moreover, it assumes that there can be no other gene between a cis-regulatory gene and a target gene and cis-regulatory regions can be associated with only one gene. Therefore, this method is expected to miss more distal cis-regulatory regions or regions regulating more than one gene. This limitation of our method can be addressed using chromosome conformation capture techniques (3C) or other adaptations as 4C, 5C, and Hi-C (Dekker et al. 2002; Simonis et al. 2006; Dostie et al. 2006; Lieberman-Aiden et al. 2009). A fundamental limitation of the rank aggregation approach that we used is the assumption that the different input weights have equal contribution to inferring interactions. However, this might not always be the case; therefore, future work needs to be performed to determine the maximum contribution for each edge. We inferred transcription factor spatial networks using the information derived from our binding and regulatory networks and genome-wide *EP300* ChIP-seq and RNA-seq. Genome-wide approaches rely on the mean signal from a bulk of tissues and cell types, which can result in missing signals stemming from heterogeneity. Using single cell or tissue-specific ChIP-seq and RNA-seq experiments will allow us to develop more detailed and comprehensive gene regulatory networks. Finally, our method was validated on a limited set of transcription factors. We believe that our method will benefit if it is tested on a more well-characterized organism, as human or mouse, with a plethora of data available.

Overall, we described a novel ensemble method to infer GRNs. Ensemble methods combine a range of information in a single model with the goal to obtain better predictions. Likewise, we saw that different data types were complementary to each other and our method improved the inferred interactions. We were able to infer stage-specific networks, which we use to study dynamics during development. We believe that our method can be applied to other organisms and provide new insights regarding gene regulation.

MATERIALS AND METHODS

Datasets

ChIP-seq

We used publicly available datasets for *EP300* ChIP-seq with the following Gene Expression Omnibus (GEO) (Barrett et al. 2013) accessions: GSM1659920 (stage 9), GSM1659921 (stage 10.5), GSM1659924 (stage 12.5), GSM1659925 (stage 12.5) and GSM1659926 (stage 30).

For TF ChIP-seq we used publicly available datasets with the following GEO accessions: GSE30146 (Smad2/3), GSM1298090 and GSM1298091 (Foxh1), GSM1180932 (Tbxt), GSM1180934 (Eomes) and GSM1867400 (beta-catenin). For Gsc and Otx ChIP-seq we used publicly available datasets for TF ChIP-seq with the following DDBJ Sequence Read Archive (DRA) (Kodama et al. 2012) accessions: DRA000576 (Gsc) and DRA000508 (Gsc).

RNA-seq

We used publicly available datasets for RNA sequencing data (RNA-seq) with the following GEO (Barrett et al. 2013) accessions: GSM1606184 and GSM1606327 (stage 8), GSM1606184 and GSM1606328 (stage 9), GSM1606190 and GSM1606334 (stage 10.5), GSM1606196 and GSM1606340 (stage 12.5), GSM1606205 and GSM1606349 (stage 16), GSM1606228 and GSM1606229 (stage 30) and GSE81458 (animal cap, vegetal mass, and dorsal, lateral and ventral marginal zones).

CIS motif database

We created a non-redundant database of TF motifs by clustering all vertebrate motifs from the CIS-BP database using GimmeMotifs (Weirauch et al. 2014; van Heeringen and Veenstra 2011).

ChIP-seq and RNA-seq analysis

ChIP-seq reads were mapped to the *X. tropicalis* genome (Xt9.0) using bwa mem (version 0.7.10-r789) with default settings (H. Li and Durbin 2009). Duplicate reads were marked using bamUtil v1.0.2. Peaks were called on the ChIP-seq data with only the uniquely mapped reads using MACS (version 2.1.0.20130306) (Y. Zhang et al. 2008) relative to the Input track using the standard settings and a q-value of 0.01. Fragment size was determined using phantompeakqualtools-2.0 (Kharchenko, Tolstorukov, and Park 2008; Landt et al. 2012).

Quantification of expression levels was performed on RNA-seq data (Owens et al. 2016; Blitz et al. 2017), using kallisto version 0.43.0 (Bray et al. 2016) with default settings and the *X. tropicalis* v9.0 assembly.

Defining cis-regulatory regions

Ep300 peaks from the five developmental stages were combined using bedtools intersect (version v.2.20.1) (Quinlan and Hall 2010). In case of overlapping (+/-25bp) summits, the confidence score from MACS2 was used to determine the strongest peak which then was included in the final dataset. Summits were extended to +/- 100bp.

Motif analysis

Cis-regulatory regions we scanned for motifs using gimme scan' from the GimmeMotifs v0.8.6 package (van Heeringen and Veenstra 2011) using the settings -b -n 1 -c 0. T. Transcription factors were linked to motifs based on the annotation for *X. tropicalis*, mouse and human from CIS-BP. This resulted in a total of 480 motifs for 942 *X. tropicalis* transcription factors.

Binding network inference

PWM score

Our first input was the database of TF motif scores from the set of 480 motifs for 942 known TFs (Weirauch et al. 2014). To correct for motif length size bias, we performed z-score normalization on the motif scores. Normalization was done per motif, based on motif matches to random genomic regions using the same motif scan settings. Z-scores were scaled from zero to one, with one being the highest and zero the lowest.

Peak intensity

Our second input data set consisted of genome-wide location ChIP-seq data for *Ep300* (Hontelez et al. 2015) across stage 9, 10.5, 12.5, 16 and 30 embryos. We calculated the RPKM (Reads Per Kilobase of transcript per Million mapped reads) (Wagner, Kin, and Lynch 2012) for each cis-regulatory region. RPKM levels were scaled from zero to one, with one being the highest and zero the lowest.

Transcription factor expression

Out third input data-set was the quantification of gene expression that was performed using kallisto version 0.43.0 (Bray et al. 2016). Expression levels (TPM) were scaled from zero to one, with one being the highest and zero the lowest.

Rank and scaling of scores

Scores were ranked using the function stats.rankdata of scipy version 1.1.0. Scores were scaled from 0 to 1 using the function preprocessing.minmax_scale of sklearn version 0.0.

Rank aggregation

To combine the PWM score, peak intensity score and transcription factor expression score we used mean rank aggregation.

Regulatory network inference

Correlation of mRNA levels

We used the RNA-seq quantification data for Owens et al. data (Owens et al. 2016), which were obtained from the authors. We calculated the Pearson correlation between all TF-gene pairs and we considered edges where the Pearson correlation score between TF and target gene was more than 0.7 and the adjusted p-value was less than 1e-5.

Influence score

Using an adaptation of the Mogrify approach described by Rackham et al. (Rackham et al. 2016), we predicted the transcription factors responsible for the differentially expressed genes between developmental stages. For each TF we built a local network, up to the third degree. Using the following equation we assigned an influence score for each node in the network, based on its distance from the TF of interest, the out-degree of the parent node and the change in expression between the stage of interest and the previous stage.

$$N_{x,n}^{S} = \sum_{r \in V_x}^{S} G_r^S \cdot \frac{1}{L_{r,n}} \cdot \frac{1}{O_{r,n}}$$

Where $r \in V_x$, is each gene (r) in the set of nodes (V_x) that make up the local subnetwork of transcription factor x. Where $L_{r,n}$, is the level (or the number of steps) that gene r is away from transcription factor x in the network n. Nodes located further from the TF had less effect on the influence score. Where $O_{r,n}$, is the out-degree of the parent of gene r in the network n. Highly ubiquitous transcription factors were prevented from getting high artificial score. Where G_r^S , is the log-transformed fold changes in expression and FDR-adjusted P values. The G_r^S was calculated using the following equation:

$$G_r^S = |L_x^S| \, (-log_{10} P_x^S)$$

Where L_x^S is the log-transformed fold change in the expression of gene x in sample s and P_x^S is the adjusted P value for gene x in sample s. To calculate differential expression between stages, we used DESeq (Anders and Huber 2010) with the default settings.

The influence score for each transcription factor was the sum of the scores from all the nodes in its local network. A node present in multiple edges is taken into account only once

and at the smallest degree (Lr,n). For example, if a gene is direct target (one step away) of a TF and a second degree target (two steps away), we count only the direct target. Self-regulating nodes were not taken into account.

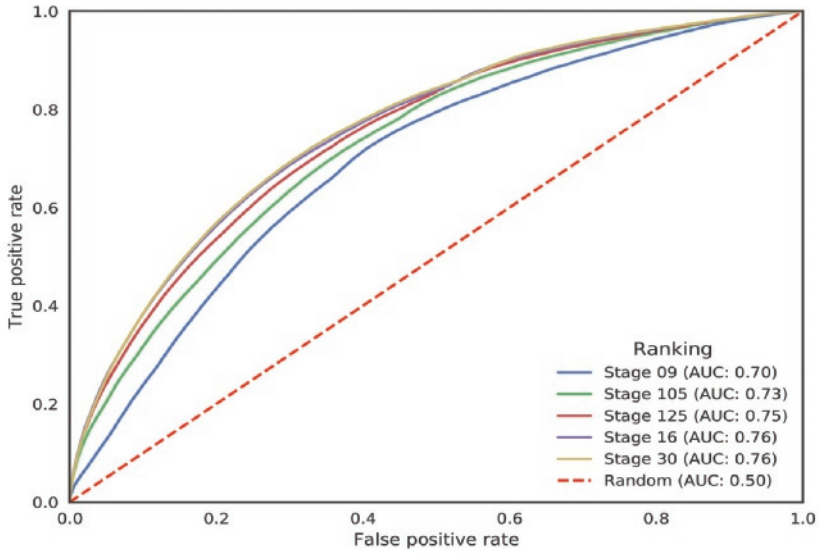
Evaluation

To evaluate the performance of our methods we used the receiver operating characteristics (ROC) area under the curve (AUC) which measures the true positive rate (TPR) against false positive rate (FPR). A random prediction will correspond to an AUC score of 0.5 and a perfect prediction to a score of 1.0. As gold standards, we used co-citation data and interactions from experimentally validated mesoderm networks. The co-citation data was downloaded from Xenbase (http://www.xenbase.org) and consists of *Xenopus* genes and TFs co-cited in the same paper. The idea and assumption behind this rationale are that genes and TFs cited together are likely to be interacting. Experimentally validated mesoderm networks were obtained from the literature (Koide, Hayata, and Cho 2005; Charney et al. 2017). Mesendoderm networks were assembled using data from X.*tropicalis* and X.*laevis*. TFs which have been shown to be essential for mesendoderm and early endoderm were included into the networks. Edges were taken into account if TFs and target genes had a strong correlation in expression changes. Based on the effect of TF on the target gene, they were required to be consistently expressed or repressed in a spatiotemporal manner. Moreover, experimental data as ChIP, EMSA, DNase footprinting and reporter gene assays were used to show if TFs and targets had direct physical interaction.

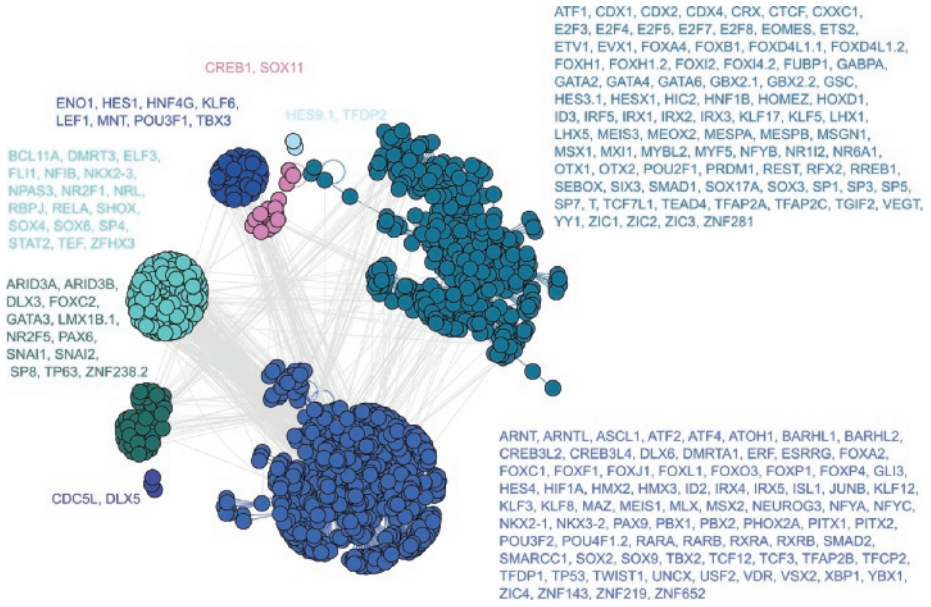
Community detection

We used the igraph library version 0.7.1.post6 in Python to identify communities. We used the community_edge_betweenness function which is based on the betweenness of the edges in the network. Directionality and weights of the edges were not taken into account. Clusters were set to None, so dendrogram was cut at the level which maximizes the modularity.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Validation of stage-specific binding networks using co-citation data. We validated the performance in predicting binding using co-citation literature data. Shown is the AUC for each of the networks. Our method performed well in all five stages with a ROC AUC between 0.70 and 0.76



Supplementary Figure 2. We examined the community structure of stage 10.5 network and identified eight distinct communities. Shown are the eight communities as identified using the edge betweenness as the centrality measure and the fast greedy modularity optimization algorithm for finding community structure. Adjacent to the communities are the gene names being part of the corresponding community. The text color matches the nodes color of each community.

REFERENCES

Albert, Réka. 2005. "Scale-Free Networks in Cell Biology." Journal of Cell Science 118 (Pt 21): 4947–57.

Alcalá-Corona, Sergio A., Tadeo E. Velázquez-Caldelas, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. 2016. "Community Structure Reveals Biologically Functional Modules in MEF2C Transcriptional Regulatory Network." Frontiers in Physiology 7 (May): 184.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." Genome Biology 11 (10): R106.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." Nature Genetics 25 (1): 25–29.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. "NCBI GEO: Archive for Functional Genomics Data Sets--Update." Nucleic Acids Research 41 (Database issue): D991–95.

Blitz, Ira L., Kitt D. Paraiso, Ilya Patrushev, William T. Y. Chiu, Ken W. Y. Cho, and Michael J. Gilchrist. 2017. "A Catalog of *Xenopus tropicalis* Transcription Factors and Their Regional Expression in the Early Gastrula Stage Embryo." Developmental Biology 426 (2): 409–17.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." Journal of Statistical Mechanics 2008 (10): P10008.

Blow, Matthew J., David J. McCulley, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2010. "ChIP-Seq Identification of Weakly Conserved Heart Enhancers." Nature Genetics 42 (9): 806–10.

Blum, Martin, and Tim Ott. 2018. "*Xenopus*: An Undervalued Model Organism to Study and Model Human Genetic Disease." Cells, Tissues, Organs, August, 1–11.

Brandes, Ulrik. 2001. "A Faster Algorithm for Betweenness Centrality." The Journal of Mathematical Sociology 25 (2): 163–77.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." Nature Biotechnology 34 (5): 525–27.

Buitrago-Delgado, Elsy, Elizabeth N. Schock, Kara Nordin, and Carole LaBonne. 2018. "A Transition from SoxB1 to SoxE Transcription Factors Is Essential for Progression from Pluripotent Blastula Cells to Neural Crest Cells." Developmental Biology 444 (2): 50–61.

Butte, A. J., and I. S. Kohane. 2000. "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 418–29.

Chan, H. M., and N. B. La Thangue. 2001. "p300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds." Journal of Cell Science 114 (Pt 13): 2363–73.

Charney, Rebekah M., Kitt D. Paraiso, Ira L. Blitz, and Ken W. Y. Cho. 2017. "A Gene Regulatory Program Controlling Early *Xenopus* Mesendoderm Formation: Network Conservation and Motifs." Seminars in Cell & Developmental Biology 66 (June): 12–24.

Chiu, William T., Rebekah Charney Le, Ira L. Blitz, Margaret B. Fish, Yi Li, Jacob Biesinger, Xiaohui Xie, and Ken W. Y. Cho. 2014. "Genome-Wide View of TGFβ/Foxh1 Regulation of the Early Mesendoderm Program." Development 141 (23): 4537–47.

Clauset, Aaron, M. E. J. Newman, and Cristopher Moore. 2004. "Finding Community Structure in Very Large Networks." Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 70 (6 Pt 2): 066111.

Clements, D., R. V. Friday, and H. R. Woodland. 1999. "Mode of Action of VegT in Mesoderm and Endoderm Formation." Development 126 (21): 4903–11.

Collart, Clara, Nick D. L. Owens, Leena Bhaw-Rosun, Brook Cooper, Elena De Domenico, Ilya Patrushev, Abdul K. Sesay, James N. Smith, James C. Smith, and Michael J. Gilchrist. 2014. "High-Resolution Analysis of Gene Activity during the *Xenopus* Mid-Blastula Transition." Development 141 (9): 1927–39.

Conlon, F. L., S. G. Sedgwick, K. M. Weston, and J. C. Smith. 1996. "Inhibition of Xbra Transcription Activation Causes Defects in Mesodermal Patterning and Reveals Autoregulation of Xbra in Dorsal Mesoderm." Development 122 (8): 2427–35.

Crease, D. J., S. Dyson, and J. B. Gurdon. 1998. "Cooperation between the Activin and Wnt Pathways in the Spatial Control of Organizer Gene Expression." Proceedings of the National Academy of Sciences of the United States of America 95 (8): 4398–4403.

Davidson, Eric H., and Michael S. Levine. 2008. "Properties of Developmental Gene Regulatory Networks." Proceedings of the National Academy of Sciences of the United States of America 105 (51): 20063–66.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." Science 295 (5558): 1306–11.

Delgado, Fernando M., and Francisco Gómez-Vela. 2018. "Computational Methods for Gene Regulatory Networks Reconstruction and Analysis: A Review." Artificial Intelligence in Medicine, November. https://doi.org/10.1016/j.artmed.2018.10.006.

Dostie, Josée, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, et al. 2006. "Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements." Genome Research 16 (10): 1299–1309.

Duncan, Anna R., and Mustafa K. Khokha. 2016. "*Xenopus* as a Model Organism for Birth Defects-Congenital Heart Disease and Heterotaxy." Seminars in Cell & Developmental Biology 51 (March): 73–79.

Eckner, R., Z. Arany, M. Ewen, W. Sellers, and D. M. Livingston. 1994. "The Adenovirus E1A-Associated 300-kD Protein Exhibits Properties of a Transcriptional Coactivator and Belongs to an Evolutionarily Conserved Family." Cold Spring Harbor Symposia on Quantitative Biology 59: 85–95.

Engleka, M. J., E. J. Craig, and D. S. Kessler. 2001. "VegT Activation of Sox17 at the Midblastula Transition Alters the Response to Nodal Signals in the Vegetal Endoderm Domain." Developmental Biology 237 (1): 159–72.

Ernst, Jason, Qasim K. Beg, Krin A. Kay, Gábor Balázsi, Zoltán N. Oltvai, and Ziv Bar-Joseph. 2008. "A Semi-Supervised Method for Predicting Transcription Factor-Gene Interactions in Escherichia Coli." PLoS Computational Biology 4 (3): e1000044.

Fortunato, Santo. 2010. "Community Detection in Graphs." Physics Reports 486 (3): 75–174.

Freeman, Linton C. 1978. "Centrality in Social Networks Conceptual Clarification." Social Networks 1 (3): 215–39.

Fukuda, Masakazu, Shuji Takahashi, Yoshikazu Haramoto, Yasuko Onuma, Yeon-Jin Kim, Chang-Yeol Yeo, Shoichi Ishiura, and Makoto Asashima. 2010. "Zygotic VegT Is Required for *Xenopus* Paraxial Mesoderm Formation and Is Regulated by Nodal Signaling and Eomesodermin." The International Journal of Developmental Biology 54 (1): 81–92.

Gao, Yan, Qing Cao, Lei Lu, Xuena Zhang, Zan Zhang, Xiaohua Dong, Wenshuang Jia, and Ying Cao. 2015. "Kruppel-like Factor Family Genes Are Expressed during *Xenopus* Embryogenesis and Involved in Germ Layer Formation and Body Axis Patterning." Developmental Dynamics: An Official Publication of the American Association of Anatomists 244 (10): 1328–46.

Gentsch, George E., Nick D. L. Owens, Stephen R. Martin, Paul Piccinelli, Tiago Faial, Matthew W. B. Trotter, Michael J. Gilchrist, and James C. Smith. 2013. "In Vivo T-Box Transcription Factor Profiling Reveals Joint Regulation of Embryonic Neuromesodermal Bipotency." Cell Reports 4 (6): 1185–96.

Girvan, M., and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." Proceedings of the National Academy of Sciences of the United States of America 99 (12): 7821–26.

Harbison, Christopher T., D. Benjamin Gordon, Tong Ihn Lee, Nicola J. Rinaldi, Kenzie D. Macisaac, Timothy W. Danford, Nancy M. Hannett, et al. 2004. "Transcriptional Regulatory Code of a Eukaryotic Genome." Nature 431 (7004): 99–104.

Heeringen, Simon J. van, and Gert Jan C. Veenstra. 2011. "GimmeMotifs: A de Novo Motif Prediction Pipeline for ChIP-Sequencing Experiments." Bioinformatics 27 (2): 270–71.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. "Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression." Nature 459 (7243): 108–12.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." Molecular Cell 38 (4): 576–89.

Hempel, Annemarie, and Michael Kühl. 2016. "A Matter of the Heart: The African Clawed Frog *Xenopus* as a Model for Studying Vertebrate Cardiogenesis and Congenital Heart Defects." Journal of Cardiovascular Development and Disease 3 (2). https://doi.org/10.3390/jcdd3020021.

Hontelez, Saartje, Ila van Kruijsbergen, Georgios Georgiou, Simon J. van Heeringen, Ozren Bogdanovic, Ryan Lister, and Gert Jan C. Veenstra. 2015. "Embryonic Transcription Is Controlled by Maternally Defined Chromatin State." Nature Communications 6 (December): 10148.

Howard, Laura, Maria Rex, Debbie Clements, and Hugh R. Woodland. 2007. "Regulation of the *Xenopus* Xsox17alpha(1) Promoter by Co-Operating VegT and Sox17 Sites." Developmental Biology 310 (2): 402–15.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. 2000. "The Large-Scale Organization of Metabolic Networks." Nature 407 (6804): 651–54.

Karimi, Kamran, Joshua D. Fortriede, Vaneet S. Lotay, Kevin A. Burns, Dong Zhou Wang, Malcom E. Fisher, Troy J. Pells, et al. 2018. "Xenbase: A Genomic, Epigenomic and Transcriptomic Model Organism Database." Nucleic Acids Research 46 (D1): D861–68.

Kharchenko, Peter V., Michael Y. Tolstorukov, and Peter J. Park. 2008. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." Nature Biotechnology 26 (12): 1351–59.

Knezevic, V., R. De Santo, and S. Mackem. 1997. "Two Novel Chick T-Box Genes Related to Mouse Brachyury Are Expressed in Different, Non-Overlapping Mesodermal Domains during Gastrulation." Development 124 (2): 411–19.

Kodama, Yuichi, Martin Shumway, Rasko Leinonen, and International Nucleotide Sequence Database Collaboration. 2012. "The Sequence Read Archive: Explosive Growth of Sequencing Data." Nucleic Acids Research 40 (Database issue): D54–56.

Kofron, M., T. Demel, J. Xanthos, J. Lohr, B. Sun, H. Sive, S. Osada, C. Wright, C. Wylie, and J. Heasman. 1999. "Mesoderm Induction in *Xenopus* Is a Zygotic Event Regulated by Maternal VegT via TGFbeta Growth Factors." Development 126 (24): 5759–70.

Koide, Tetsuya, Tadayoshi Hayata, and Ken W. Y. Cho. 2005. "*Xenopus* as a Model System to Study Transcriptional Regulatory Networks." Proceedings of the National Academy of Sciences of the United States of America 102 (14): 4943–48.

Lancichinetti, Andrea, and Santo Fortunato. 2009. "Community Detection Algorithms: A Comparative Analysis." Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 80 (5 Pt 2): 056117.

Lancichinetti, Andrea, Filippo Radicchi, José J. Ramasco, and Santo Fortunato. 2011. "Finding Statistically Significant Communities in Networks." PloS One 6 (4): e18961.

Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." Genome Research 22 (9): 1813–31.

Lee, Wei-Po, and Wen-Shyong Tzou. 2009. "Computational Methods for Discovering Gene Networks from Expression Data." Briefings in Bioinformatics 10 (4): 408–23.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science 326 (5950): 289–93.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." Bioinformatics 25 (14): 1754–60.

Liu, Zhi-Ping. 2015. "Reverse Engineering of Genome-Wide Gene Regulatory Networks from Gene Expression Data." Current Genomics 16 (1): 3–22.

Li, Wei, and Robert A. Cornell. 2007. "Redundant Activities of Tfap2a and Tfap2c Are Required for Neural Crest Induction and Development of Other Non-Neural Ectoderm Derivatives in Zebrafish Embryos." Developmental Biology 304 (1): 338–54.

Li, Xiao-Yong, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A. Pollard, Venky N. Iyer, Aaron Hechmer, et al. 2008. "Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm." PLoS Biology 6 (2): e27.

Long, Hannah K., Sara L. Prescott, and Joanna Wysocka. 2016. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." Cell 167 (5): 1170–87.

Loose, Matthew, and Roger Patient. 2004. "A Genetic Regulatory Network for *Xenopus* Mesendoderm Formation." Developmental Biology 271 (2): 467–78.

Luo, Ting, Young-Hoon Lee, Jean-Pierre Saint-Jeannet, and Thomas D. Sargent. 2003. "Induction of Neural Crest in *Xenopus* by Transcription Factor AP2alpha." Proceedings of the National Academy of Sciences of the United States of America 100 (2): 532–37.

Luo, Ting, Mami Matsuo-Takasaki, Megan L. Thomas, Daniel L. Weeks, and Thomas D. Sargent. 2002. "Transcription Factor AP-2 Is an Essential and Direct Regulator of Epidermal Development in *Xenopus*." Developmental Biology 245 (1): 136–44.

MacDonald, Bryan T., Keiko Tamai, and Xi He. 2009. "Wnt/beta-Catenin Signaling: Components, Mechanisms, and Diseases." Developmental Cell 17 (1): 9–26.

Marbach, Daniel, James C. Costello, Robert Küffner, Nicci Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, et al. 2012. "Wisdom of Crowds for Robust Gene Network Inference." Nature Methods 9 (8): 796–804.

Marbach, Daniel, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher a. Bristow, and Manolis Kellis. 2012. "Predictive Regulatory Models in Drosophila Melanogaster by Integrative Inference of Transcriptional Networks." Genome Research 22 (7): 1334–49.

Margolin, Adam A., Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context." BMC Bioinformatics 7 Suppl 1 (March): S7.

Matsui, T., J. Segall, P. A. Weil, and R. G. Roeder. 1980. "Multiple Factors Required for Accurate Initiation of Transcription by Purified RNA Polymerase II." The Journal of Biological Chemistry 255 (24): 11992–96.

Matsuo-Takasaki, Mami, Michiru Matsumura, and Yoshiki Sasai. 2005. "An Essential Role of *Xenopus* Foxi1a for Ventral Specification of the Cephalic Ectoderm during Gastrulation." Development 132 (17): 3885–94.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." Nature Biotechnology 28 (5): 495–501.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." Science 298 (5594): 824–27.

Morris, Samantha A., Patrick Cahan, Hu Li, Anna M. Zhao, Adrianna K. San Roman, Ramesh A. Shivdasani, James J. Collins, and George Q. Daley. 2014. "Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet." Cell 158 (4): 889–902.

Murgan, Sabrina, Aitana Manuela Castro Colabianchi, Renato José Monti, Laura Elena Boyadjián López, Cecilia E. Aguirre, Ernesto González Stivala, Andrés E. Carrasco, and Silvia L. López. 2014. "FoxA4 Favours Notochord Formation by Inhibiting Contiguous Mesodermal Fates and Restricts Anterior Neural Development in *Xenopus* Embryos." PloS One 9 (10): e110559.

Nakamura, Yukio, Eduardo de Paiva Alves, Gert Jan C. Veenstra, and Stefan Hoppler. 2016. "Tissue- and Stage-Specific Wnt Target Gene Expression Is Controlled Subsequent to β-Catenin Recruitment to Cis-Regulatory Modules." Development 143 (11): 1914–25.

Nieuwkoop, Pieter Dirk, and Jacob Faber. 1994. Normal Table of *Xenopus laevis* (Daudin): A Systematical and Chronological Survey of the Development from the Fertilized Egg Till the End of Metamorphosis. Garland Pub.

Obayashi, Takeshi, and Kengo Kinoshita. 2009. "Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression." DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes 16 (5): 249–60.

Oliveri, Paola, and Eric H. Davidson. 2007. "Development. Built to Run, Not Fail." Science 315 (5818): 1510–11.

Owens, Nick D. L., Ira L. Blitz, Maura A. Lane, Ilya Patrushev, John D. Overton, Michael J. Gilchrist, Ken W. Y. Cho, and Mustafa K. Khokha. 2016. "Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development." Cell Reports 14 (3): 632–47.

Parker, Stephen C. J., Michael L. Stitzel, D. Leland Taylor, Jose Miguel Orozco, Michael R. Erdos, Jennifer A. Akiyama, Kelly Lammerts van Bueren, et al. 2013. "Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants." Proceedings of the National Academy of Sciences of the United States of America 110 (44): 17921–26.

Peng, Gang, and Monte Westerfield. 2006. "Lhx5 Promotes Forebrain Development and Activates Transcription of Secreted Wnt Antagonists." Development 133 (16): 3191–3200.

Prill, Robert J., Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. 2010. "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges." PloS One 5 (2): e9202.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." Bioinformatics 26 (6): 841–42.

Rackham, Owen J. L., Jaber Firas, Hai Fang, Matt E. Oates, Melissa L. Holmes, Anja S. Knaupp, FANTOM Consortium, et al. 2016. "A Predictive Computational Framework for Direct Reprogramming between Human Cell Types." Nature Genetics 48 (3): 331–35.

Rosvall, Martin, and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." Proceedings of the National Academy of Sciences of the United States of America 105 (4): 1118–23.

Ryan, K., N. Garrett, A. Mitchell, and J. B. Gurdon. 1996. "Eomesodermin, a Key Early Gene in *Xenopus* Mesoderm Differentiation." Cell 87 (6): 989–1000.

Schmitt, Stefan M., Mazhar Gull, and André W. Brändli. 2014. "Engineering *Xenopus* Embryos for Phenotypic Drug Discovery Screening." Advanced Drug Delivery Reviews 69-70 (April): 225–46.

Segerdell, Erik, Virgilio G. Ponferrada, Christina James-Zorn, Kevin A. Burns, Joshua D. Fortriede, Wasila M. Dahdul, Peter D. Vize, and Aaron M. Zorn. 2013. "Enhanced XAO: The Ontology of *Xenopus* Anatomy and Development Underpins More Accurate Annotation of Gene Expression and Queries on Xenbase." Journal of Biomedical Semantics 4 (1): 31.

Serin, Elise A. R., Harm Nijveen, Henk W. M. Hilhorst, and Wilco Ligterink. 2016. "Learning from Co-Expression Networks: Possibilities and Challenges." Frontiers in Plant Science 7 (April): 1–18.

Shen-Orr, Shai S., Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. "Network Motifs in the Transcriptional Regulation Network of Escherichia Coli." Nature Genetics 31 (1): 64–68.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. "Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture-on-Chip (4C)." Nature Genetics 38 (11): 1348–54.

Spirin, Victor, and Leonid A. Mirny. 2003. "Protein Complexes and Functional Modules in Molecular Networks." Proceedings of the National Academy of Sciences of the United States of America 100 (21): 12123–28.

Spitz, François, and Eileen E. M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." Nature Reviews. Genetics 13 (9): 613–26.

Stender, J. D., K. Kim, T. H. Charn, B. Komm, K. C. N. Chang, W. L. Kraus, C. Benner, C. K. Glass, and B. S. Katzenellenbogen. 2010. "Genome-Wide Analysis of Estrogen Receptor DNA Binding and Tethering Mechanisms Identifies Runx1 as a Novel Tethering Factor in Receptor-Mediated Transcriptional Activation." Molecular and Cellular Biology 30 (16): 3943–55.

The Gene Ontology Consortium. 2017. "Expansion of the Gene Ontology Knowledgebase and Resources." Nucleic Acids Research 45 (D1): D331–38.

Toyama, R., P. E. Curtiss, H. Otani, M. Kimura, I. B. Dawid, and M. Taira. 1995. "The LIM Class Homeobox Gene lim5: Implied Role in CNS Patterning in *Xenopus* and Zebrafish." Developmental Biology 170 (2): 583–93.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. "ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers." Nature 457 (7231): 854–58.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. "Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent among Samples." Theory in Biosciences = Theorie in Den Biowissenschaften 131 (4): 281–85.

Wang, Shu-Ping, Zhanyun Tang, Chun-Wei Chen, Miho Shimada, Richard P. Koche, Lan-Hsin Wang, Tomoyoshi Nakadai, et al. 2017. "A UTX-MLL4-p300 Transcriptional Regulatory Network Coordinately Shapes Active Enhancer Landscapes for Eliciting Transcription." Molecular Cell 67 (2): 308–21.e6.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity." Cell 158 (6): 1431–43.

Wheeler, Grant N., and André W. Brändli. 2009. "Simple Vertebrate Models for Chemical Genetics and Drug Discovery Screens: Lessons from Zebrafish and *Xenopus*." Developmental Dynamics: An Official Publication of the American Association of Anatomists 238 (6): 1287–1308.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." Cell 153 (2): 307–19.

Xanthos, J. B., M. Kofron, C. Wylie, and J. Heasman. 2001. "Maternal VegT Is the Initiator of a Molecular Network Specifying Endoderm in *Xenopus laevis*." Development 128 (2): 167–80.

Yasuoka, Yuuri, Yutaka Suzuki, Shuji Takahashi, Haruka Someya, Norihiro Sudou, Yoshikazu Haramoto, Ken W. Cho, Makoto Asashima, Sumio Sugano, and Masanori Taira. 2014. "Occupancy of Tissue-Specific Cis-Regulatory Modules by Otx2 and TLE/Groucho for Embryonic Head Specification." Nature Communications 5 (January): 4322.

Yoon, Se-Jin, Andrea E. Wills, Edward Chuong, Rakhi Gupta, and Julie C. Baker. 2011. "HEB and E2A Function as SMAD/FOXH1 Cofactors." Genes & Development 25 (15): 1654–61.

Zawel, L., and D. Reinberg. 1993. "Initiation of Transcription by RNA Polymerase II: A Multi-Step Process." Progress in Nucleic Acid Research and Molecular Biology 44: 67–108.

Zhang, J., D. W. Houston, M. L. King, C. Payne, C. Wylie, and J. Heasman. 1998. "The Role of Maternal VegT in Establishing the Primary Germ Layers in *Xenopus* Embryos." Cell 94 (4): 515–24.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." Genome Biology 9 (9): R137.

CHAPTER SIX

Discussion

Gene regulation is the mechanism that controls the activation and repression of genes in the genome. Different cell types, containing the same DNA, can have a different set of genes expressed at the same time. The precise spatial and temporal control of gene regulation is an essential part of development and other vital processes. Chromatin accessibility, histone modifications, and protein-coding genes are some of the factors that control when and where a gene will be activated.

In this thesis, we studied the dynamic process of gene regulation during embryonic development and in the context of evolution. Using high-throughput sequencing technology we explored the developmental origins of epigenetic regulation and chromatin dynamics (Chapter Two). We described a software package that allows for simple exploration, clustering, and visualization of high-throughput sequencing data mapped to a reference genome (Chapter Three). Next, we looked into the consequences after the interspecific hybridization genome duplication in *Xenopus laevis* (Chapter Four). Finally, we studied the interplay of transcription factors with their target genes from a gene regulatory networks perspective (Chapter Five).

1 REGULATORY DYNAMICS DURING EMBRYONIC DEVELOPMENT

1.1 Chromatin dynamics during embryonic development

In Chapter Two, we looked at chromatin dynamics during the development of *X. tropicalis* embryos. We performed ChIP-seq of eight histone modifications, RNAPII and the coactivator p300 at five stages of development. We showed that the deposition of H3K4me3 and the Polycomb-repressive H3K27me3 modifications are largely maternally defined and they are deposited hours before transcription activation on regions with hypomethylated DNA. In contrast, the recruitment of the H3K27 acetyltransferase *Ep300* to distal regulatory elements is mainly under the control of zygotic factors.

Using a hidden Markov model approach, we identified seven groups of histone modifications and bound proteins; Polycomb (H3K27me3, deposited by Polycomb Repressive Complex 2 (PRC2)), poised enhancers and promoters, active enhancers, transcribed genes, promoters, heterochromatin and unmodified. Using alluvial plots of state coverage per stage, we showed the transitions of each group across the five developmental stages. We saw that all groups increase in coverage during development, except for the unmodified. This suggests that during development the epigenome starts as unmodified and acquires modifications during later development. Promoter coverage remains relatively constant during development, while Polycomb, transcribed, promoter, heterochromatin states show an increase in coverage. Enhancers are the most dynamic group during development. Despite the decrease in coverage,

the unmodified regions represent the largest group at the tailbud stage , with 67% of the total epigenome remaining naive. Here a naive epigenome refers to a state not primed for modifications and bound proteins. Global enrichment levels of modified regions had similar dynamics, which confirms that the epigenome starts as unmodified and chromatin marks emerge during development. Promoter marks were the first to be specified at or before the blastula stage, followed by enhancer and heterochromatic marks during late blastula and gastrulation stages. We saw that the promoter marks H3K4me3 and H3K9ac emerge and increase in coverage prior to the start of embryonic transcription. Previous studies from our group reported similar results with marking of H3K4me3 preceding the mid-blastula transition (Akkers et al. 2009). In addition, we showed that the promoter-permissive H3K4me3 mark and the Polycomb-repressive H3K27me3 modification were maternally defined and independent of zygotic transcription. A large number of neurula and tailbud stages promoters are maternally defined, which shows that maternal epigenetic control extends post gastrulation. By contrast, p300-bound regions were zygotic defined. Genes with maternally defined H3K4me3 have fewer enhancers associated with them, while the rest required could require zygotic defined enhancers. Activation of zygotically defined enhancers involved binding of transcription factors that can open chromatin, recruit coactivators or establish looping interactions with promoters. These results show the combinatorial action of maternal and zygotic factors and their effect on chromatin state. Maternal control is maintained during development in proximal promoter elements, while the zygote controls distal enhancers.

The approach we used in Chapter Two to identify combinations of chromatin marks and their dynamics is based on genome segmentation and binarization (Ernst and Kellis 2012). The genome is divided in 200-nucleotide intervals and models the presence or absence of each chromatin mark. In Chapter Three, we demonstrated a different method to identify dynamic patterns between different conditions or developmental stages using signal intensity. Most of the methods used to cluster genomic regions use a binning approach followed by clustering using the Euclidean distance. This is sufficient for clustering regions on basis of the spatial patterns relative to the region of interest, but disregards dynamic clusters. The demonstrated method allows the identification of dynamic clusters of genomic regions based on a single value derived from the number of reads in the feature's center. The regions are clustered on basis of the Pearson correlation of read counts, which allows the identification of dynamic clusters. We applied this method to DNase I hypersensitive sites in H1 human embryonic stem cells differentiated into mesenchymal, mesendoderm, neuronal progenitor and trophoblast lineages and successfully identified the dynamic sites specific to those lineages as they were described by Xie et al. (Xie et al. 2013). In comparison to the hidden Markov model approach, our method focuses only on regions of interest, e.g., peaks called on ChIP-seq experiments, and

not on binned intervals in the whole genome. However, depending on the number of peaks and marks used, this may be computationally demanding.

1.2 Dynamics at regulatory elements during embryonic development

In Chapter Two, we saw that p300-bound enhancers were the most dynamic element during the development of *X. tropicalis*. By modeling transcription factor motif contributions to p300 binding sites across multiple developmental stages, we found enriched transcription factor motifs. These results suggest that the enriched transcription factors recruit p300 in a stage-specific manner. Studies in human showed similar results with enhancers being the most dynamic element of the genome (Heintzman et al. 2009; Maston et al. 2012; Buecker and Wysocka 2012). Experiments in cervical carcinoma HeLa, immortalized lymphoblast GM06690, leukemia K562, embryonic stem cells, and BMP4-induced ES cells showed that majority of enhancers were cell-type specific (Heintzman et al. 2009). Similar results have been shown in hematopoietic cell types where chromatin accessible regions, which may correspond to enhancers, were more cell-type specific than expression patterns (Corces et al. 2016). These results highlight the importance of enhancers in controlling gene expression in a cell type-specific manner. However, enhancers can be shared among cell types or conditions. In *X. tropicalis*, we expect to have a similar diversity of enhancers in different cell types which highlights the importance of studying gene regulation in a more spatial and temporal manner. Enhancer elements increased in numbers during development which subsequently lead to the formation of large enhancer clusters, also referred to as super-enhancers (Parker et al. 2013; Whyte et al. 2013). These clusters were formed by seeding of individual p300-bound enhancers. The majority of enhancer clusters increased in genomic coverage during development by newly gained p300 binding at enhancers. Future work needs to be performed to address to which extent seeding causes the relaxation and opening of the local chromatin and activity of neighboring enhancers. We saw that the formation of enhancer clusters depends on embryonic transcription with around half of the zygotically defined p300-bound regions contributing to the enhancer clusters. On the other hand, only one third of the maternally defined p300-bound regions contributed to the enhancer clusters. Enhancer clusters were developmental stage-specific. Other studies reported similar results with enhancer clusters being implicated in cell differentiation and associated with genes coding for developmental regulators (Whyte et al. 2013; Hnisz et al. 2013; Pott and Lieb 2015). The maternally defined p300 was mostly recruited to promoter-proximal regions enriched with promoter-related motifs. In contrast, zygotically defined p300 was predominantly recruited to enhancer regions decorated with H3K4me1 in the absence of the promoter-associated mark H3K4me3. Enhancers can contain binding sites for pioneer transcription factors. Pioneer transcription factors are responsible for opening up chromatin, recruitment of co-activators and establishing looping interactions with promoters (Zaret and Carroll 2011; Iwafuchi-Doi and Zaret 2014; Cockerill 2011; Lupien et

al. 2008). We and others saw that both, maternally and zygotically defined, enhancer groups recruit embryonically regulated transcription factors (Gentsch et al. 2013; Chiu et al. 2014). However, future experiments have to be performed to examine the regulatory dynamics of the maternal-to-zygotic transition and the combinatorial interplay of maternal and zygotic factors.

In Chapter Five, we looked into regulation dynamics from a gene regulatory network perspective. We saw that different sets of transcription factors drive the regulatory program for each developmental stage. Using computationally predicted gene regulatory networks, we scored the transcription factors based on their influence on the expression of their target genes. We grouped the transcription factors in seven clusters with distinct patterns in the influence score during embryonic development. We saw transcription factors having high influence in either early or late development, as well factors that are predicted to be influential across five developmental stages. Similar results were reported by others, with specific transcription factors being responsible for regulating cell fate and driving expression in different developmental stages (Tiwari et al. 2018; P. Huang et al. 2011; Ieda et al. 2010; Sekiya and Suzuki 2011; Vierbuchen et al. 2010; Yu et al. 2013; Godoy et al. 2015).

For the transition from blastula to gastrula stage (Nieuwkoop-Faber stage 9 to stage 10.5), we predicted Tcf7l1, Otx2, Sox17a, Tbxt, Sox2 and Vegt as some of the top influential transcription factors. In *Xenopus*, these transcription factors start increasing in expression before gastrula stage and are known to be necessary for gastrulation (Owens et al. 2016; Hoffman, Wu, and Merrill 2013; Acampora et al. 1995; Lolas et al. 2014; Kishi et al. 2000; Horb and Thomsen 1997). During *Xenopus* development, Tcf7l1 promotes the transcription of Klf4, a gene known to be crucial for germ-layer differentiation and body axis patterning (Cao et al. 2017, 2012). Moreover, it acts as a mediator of Wnt signaling by forming a complex with β-catenin and regulates pattern dorsal-ventral axis specification (Molenaar et al. 1996; Brannon et al. 1997). Likewise, during mouse development, it has been shown that Tcf7l1 is essential for the specification of mesoderm by coordinating lineage specification during gastrulation (Hoffman, Wu, and Merrill 2013). In *Xenopus*, the T-box transcription factor Vegt is essential for endoderm formation and controls the primary germ layer specification in *Xenopus* embryos (Jian Zhang et al. 1998). It regulates endodermal genes, such as Sox17a, and anterior endodermal genes (Xanthos et al. 2001). Afterward, using the Nieuwkoop-Faber stage 10.5 network, we looked into spatial regulatory dynamics for the animal cap, vegetal mass, lateral marginal, dorsal marginal and ventral marginal zones. We constructed spatially resolved TF-TF regulatory networks and identified the transcription factors essential for gene regulation in these zones. For the animal cap, we identified a dense subnetwork is formed by Klf17, Foxi4.2, Tfap2a, Tfap2c and Lhx5. Foxi4.2 is required for correct ventral specification of the early head ectoderm (Matsuo-Takasaki, Matsumura, and Sasai 2005). In *Xenopus*, the AP2alpha

transcription factors Tfap2a and Tfap2c are known to target epidermal, as well as neural crest genes (Luo et al. 2002, 2003). The Kruppel-like factor Klf17 is both maternally and zygotically expressed and is enriched at the animal pole in early embryos, however, its role in early *Xenopus* development is still known (Gao et al. 2015). Lhx5 expressed in the entire ectoderm and is likely involved in the development of the nervous system (Peng and Westerfield 2006; Toyama et al. 1995). Based on the function of the known function of the genes in this animal cap network, we hypothesize that these transcription factors are important in the specification of neuronal versus non-neuronal ectoderm. In the ventral, lateral and dorsal marginal zone networks, we found the T-box transcription factors, Vegt, Eomes, and Tbxt, to have an essential role. The role and importance of these T-box transcription factors in mesoderm formation and differentiation has already been well described in the literature (Kofron et al. 1999; J. Zhang et al. 1998; Ryan et al. 1996; Fukuda et al. 2010; Knezevic, De Santo, and Mackem 1997; Conlon et al. 1996). The zygotically expressed transcription factor Foxa4 was present in all marginal zone networks. Foxa4 is known to cooperate with T-box transcription factors in dorsal mesoderm formation (Murgan et al. 2014). In the ventral marginal zone network, the transcription factors Vegt and Sox17a were among the most important transcription factors. This finding is confirmed by the literature where Vegt and Sox17a are known to be required for embryonic endoderm development (Howard et al. 2007; Engleka, Craig, and Kessler 2001). These results highlight that transcription factors not only are expressed in a time-specific manner but also spatial-specific.

Concerning regulatory dynamics during development, I believe that future research should focus on spatio-temporal gene regulation. As discussed previously, enhancers dynamically control gene expression in a cell type-specific manner. Likewise, transcription factors are expressed in a time- and spatial-specific manner. Our network analyses were based on experiments performed on whole embryos and results could be affected by cell-type variations. I suggest that in the future networks could be inferred using data from specific timepoints and parts of the embryo. Studies in pluripotent stem cells at single-cell level have shown that there is variability among different populations (S. Huang 2009). Nowadays, with the emergence of single-cell technologies we can gather data from individual cells and get more nuanced measurements. Single-cell experiments will aid the inference of germ-layer-specific networks and will allow unprecedented opportunities to study the cellular heterogeneity and the underlying regulatory programmes. It will be interesting to see the interplay of transcription factors with their target genes at different parts of the embryo during development.

1.3 Regulatory dynamics in the context of evolution

In Chapter Four, we examined enhancers patterns in the context of the evolution of *X. laevis* after the interspecific hybridization genome duplication. X.*laevis* resulted from the hybridization of two closely related species about 17 million years ago and its genome

consists of two subgenomes. The two genomes are referred to as L (long chromosomes) and S (short chromosomes). We saw that p300 recruitment was remarkably different between L and S loci, with differences in both p300 peak intensity and the number of peak regions across homologous loci. Only 13% of the p300-bound enhancers were conserved among the two genomes, while 40% of promoters were found to be conserved. Our findings coincide with studies in vertebrates that enhancers are less conserved and evolve considerably faster than promoters and are therefore more dynamic during evolution (Blow et al. 2010; Hsu and Ovcharenko 2013; Villar et al. 2015). Studies in heart enhancers showed that only a small number was conserved between human and mouse (Hsu and Ovcharenko 2013). Likewise, liver enhancers were rarely conserved and evolved faster within mammalian genomes (Villar et al. 2015). However, although they are not common, highly conserved enhancers tend to be near genes important for fundamental processes, such as embryonic development (Boffelli, Nobrega, and Rubin 2004; Woolfe et al. 2005). Despite the low conservation of enhancers, transcription factor binding sites are usually conserved among closely related species (Schmidt et al. 2010; Arnold et al. 2014).

2 THE ROLE OF REPETITIVE ELEMENTS IN REGULATORY REMODELING

Transposable elements were discovered in the 1950s by Barbara McClintock through her pioneering work in cytogenetic analyses of maize chromosomes (McClintock 1950). Since then these elements have been detected in more plants and animals (Kazazian 2004). In mouse and human, around 50% of the genome is derived from transposable elements (Mouse Genome Sequencing Consortium et al. 2002; Lander et al. 2001; Bannert and Kurth 2004).

Initially, transposable elements were considered as parasitic or junk DNA and harmful for the organisms. However, in Chapter Four we saw that a small fraction of those transposable elements could be evolutionarily beneficial for organisms. Transposable elements have the ability to control gene expression and modify genomic architecture by introducing new regulatory regions and transcription factor binding sites. Others reported similar results with transposable elements being responsible for new cis-regulatory elements and contributing to a large number of transcription factor binding sites (Friedli and Trono 2015; Bourque et al. 2008; Sundaram et al. 2014). In Chapter Four, we looked at the conservation of enhancers compared to *X. tropicalis*. We found 1,214 and 1,237 enhancers, in L and S genomes respectively, lacking any conservation with either the other subgenome or *X. tropicalis*. We looked into those "new" enhancers and we saw that they were enriched with three repeat annotations, named REM1, Kolobok-T2, and family-131. These repeats carried binding sites for the transcription factors Plag1, Eomes/Tbx21 T-box factors, Sox18, Mecom/Prdm16, and the Six3/Six6 homeobox

factors. Looking at the immediate effects of hybridization in *X. tropicalis* × *X. laevis* hybrid embryos, we found 629 new enhancers in the *X. tropicalis* genome, while none was found in *X. laevis*. Looking into these newly gained enhancers, we found that they overlapped with three repeat annotations (family - 451, 203, and 189). These three repeats had 80% similarity with the PiggyBac-N2A DNA transposons repeats. Upon further examination, we found that these repeats contained transcription factor binding sites for Homeodomain and T-box binding factors. In addition, we saw that the newly gained enhancers were significantly enriched with the heterochromatic mark H3K9me3 in normal *X. tropicalis* embryos. Recent studies from our group found that young DNA transposons are enriched with H3K9me3 (van Kruijsbergen et al. 2017). These findings suggest that these are young DNA transposable elements, proliferated after the split with *X. tropicalis* or derepressed in the *X. laevis* egg and contribute sequence variation to the genome.

Our results are in line with studies in mouse and human that demonstrate that transposable elements facilitate newly gained genome-specific p300 peaks and binding sites (Bourque et al. 2008; Sundaram et al. 2014). These results highlight that transposable elements can contribute to the evolutionary dynamics of species.

3 GENE REGULATORY NETWORKS

3.1 Inferring Gene Regulatory Networks

A main feature of enhancers is their ability to function as a binding platform for multiple transcription factors. Transcription factors can play different roles in opening up chromatin, recruitment of co-activators and establishing looping interactions with promoters. They recognize and bind to specific motifs in the vicinity of their target genes. All interactions between transcription factors and target genes form a gene regulatory network which orchestrates the transcriptional regulation. Elucidating gene regulatory networks will have a significant impact on biology and medicine, with applications ranging from gaining new insights into regulatory mechanisms, to personalized medicine and identification of potential new drug targets (Bower and Bolouri 2004; Blais and Dynlacht 2005; T. I. Lee et al. 2002; Ghosh and Basu 2012; Fortney et al. 2013; Madhamshettiwar et al. 2012). In Chapter Five, we described a novel ensemble method to infer gene regulatory networks by integrating the binding of the p300 coactivator, transcription factor expression, and transcription factor motifs.

Our method infers binding networks by predicting the binding of transcription factors in cis-regulatory regions. The value of such an approach is that we bypass the impossible task of performing ChIP-seq for all the TFs at all stages or tissues. However, binding of a transcription factor in a gene locus does not always imply regulation. Co-expression data are widely used to

infer gene regulatory networks by computing a similarity measurement between gene pairs, such as correlation coefficient (Langfelder and Horvath 2008; Obayashi and Kinoshita 2009; Opgen-Rhein and Strimmer 2007). If the measurement is greater than a specified threshold, then the pair is connected in an undirected manner. Based on the premise that genes with similar expression can be functionally related we introduced co-expression data to infer directed gene regulatory networks.

We and others observed that combining predictions from multiple inference methods can improve the inference of a gene regulatory network (Marbach, Costello, et al. 2012). A study in *Drosophila* inferred a gene regulatory network by combining data from ChIP-seq experiments, conservation from 12 *Drosophila* species, gene expression and histone modifications. They showed that the combination of different sources improved the prediction (Marbach, Roy, et al. 2012). The principle of using an ensemble of predictors appears to be a powerful approach in constructing gene regulatory networks.

Currently, our approach is based on unsupervised learning using simple rank aggregation. For each edge, the different predictors have the same impact on the consensus network. However, this might not be the ideal approach because data do not necessarily have the same biological relevance. Therefore, we could look into a different rank aggregation approach in the future. Computational techniques and algorithms have been proposed for comparison and integration of different classes of information using different hypotheses, such as Bayesian-based reasoning and order statistics (Badgeley, Sealfon, and Chikina 2015; Weile et al. 2012; I. Lee et al. 2004; Kolde et al. 2012; Stuart et al. 2003). From preliminary analyses we saw that introducing regression analysis and training data does improve the predictions. Currently, our ensemble approach is unsupervised. I believe it could benefit by making it supervised, or semi-supervised, and train it on several transcription factor datasets. The limited availability of TF ChIP-seq data in *Xenopus* made it not feasible to achieve this for the work presented in Chapter Five. For future work, I would recommend implementing and testing the approach in human or mouse, for which an abundance of data is publicly available. Likewise, our method was validated on a small set of experimentally validated interactions and transcription factor ChIP-seq experiments. Having available more training and validation data may improve the predictions and strengthen our approach. In Chapter Five we used whole embryo data, where measurements are based on the population average. As mentioned above, future research should focus on single-cell experiments. Many approaches and studies have emerged that put efforts in inferring and studying GRNs through single-cell data (Aibar et al. 2017; Matsumoto et al. 2017; Herbach et al. 2017; Pina et al. 2015; Patel et al. 2014). Finally, a last limitation of our approach is that it can be computationally demanding. Storing, incorporating and analyzing a large amount of data can be cumbersome and an impossible task for a simple workstation.

4 BIOINFORMATICS CHALLENGES AND ADVANCEMENTS IN THE ERA OF BIG DATA

The rise of high-throughput sequencing technologies during the last decade has led to an enormous amount of data. New technologies can sequence millions of reads in parallel and in one week they can produce hundreds of gigabytes of data from a single machine. To illustrate, the European Bioinformatics Institute (EBI), one of the world's largest repositories of biological data, had a storage capacity of 120 Petabytes in 2017. This is expected to grow by 40-50% each year (Chen and Gao 2016; Cook et al. 2018). This accelerated growth led us to the "big data" era and has raised the challenges of storing, sharing and analyzing these data (Peek, Holmes, and Sun 2014; Marx 2013).

As the cost of sequencing continues to drop, organizations are facing a significant challenge in handling all the data. The cost of storage and the time needed to analyze data can be a bottleneck for small organizations. This has led many towards cloud services for storage and computation (Kashyap et al. 2015). Cloud services provide an attractive and affordable solution for handling data. Resources, such as processors and memory, are dynamically scalable and available on demand. The costs for pay-as-you-go services can be significantly lower than purchasing, maintaining and supporting in-house infrastructure. Cloud services can be purchased in three main models; Infrastructure as a service (IaaS), Platform as a service (PaaS) and Software as a service (SaaS). In IaaS the provider provides access to computational resources and organizations can deploy their own platforms. In PaaS, organizations can deploy and manage their own applications, whereas in Saas the provider provides access to their own applications.

Nowadays, cloud providers, such as Amazon Web Services, Microsoft Azure, Google Cloud Platform (GCP) and IBM Watson, are actively advertising the use of their services in genomic research ("Genomics Cloud Computing", "Microsoft Genomics | Microsoft Azure", "Google Genomics - Store, Process, Explore and Share | Cloud Genomics | Google Cloud", "IBM Watson for Genomics - Overview - United States"). However, adoption of cloud computing is not a simple task and comes with its drawbacks and concerns among the community. These include reliability, security, privacy, compatibility, and ownership of the data (Tripathi et al. 2016). In fact, with the 2018's General Data Protection Regulation (GDPR) from the European Union, organizations face major challenges and legal responsibilities in managing sensitive information, such as clinical data.

In conclusion, cloud services are an attractive solution in the field of genomics. They can offer affordable storage, memory, computation power and services to organizations with a

single click of a button. In the era where data are getting bigger and bigger, cloud computing is a particularly attractive option and it seems it is here to stay. However, adapting to this "new" trend is a rather complicated task that still requires a lot of research.

5 CONCLUDING REMARKS

The process of embryonic development is tightly controlled by gene regulation, which is coordinated by TFs interacting with the DNA and the surrounding chromatin environment. TFs bind to cis-regulatory elements, proximal or distal to the transcription start sites, and influence transcription. In the last decade, advancements in the field of sequencing made it possible to identify and study the complete cellular environment in different cell types, conditions or developmental stages. Modern techniques, such as single-cell genomics and CRISPR-Cas9-based techniques, can advance the research in the field. Single-cell techniques on germ-layer specific experiments can yield spatially resolved measurements, while CRISPR-Cas9 can be used to target complexes and activate or repress regulatory regions. The field of bioinformatics needs to advance at a similar rate. The large amount of data requires new computational methods, new algorithms for faster and more efficient data processing and new data storing solutions. New advancements will make the integration of the large volumes of data feasible. This will help to decipher and model the complex gene regulatory networks that control critical biological processes and unravel new interactions. These will aid in the development of new treatments against diseases and have an impact in the area of regenerative medicine.

REFERENCES

Acampora, D., S. Mazan, Y. Lallemand, V. Avantaggiato, M. Maury, A. Simeone, and P. Brûlet. 1995. "Forebrain and Midbrain Regions Are Deleted in Otx2-/- Mutants due to a Defective Anterior Neuroectoderm Specification during Gastrulation." Development 121 (10): 3279–90.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. 2017. "SCENIC: Single-Cell Regulatory Network Inference and Clustering." Nature Methods 14 (11): 1083–86.

Akkers, Robert C., Simon J. van Heeringen, Ulrike G. Jacobi, Eva M. Janssen-Megens, Kees-Jan Françoijs, Hendrik G. Stunnenberg, and Gert Jan C. Veenstra. 2009. "A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in *Xenopus* Embryos." Developmental Cell 17 (3): 425–34.

Arnold, Cosmas D., Daniel Gerlach, Daniel Spies, Jessica A. Matts, Yuliya A. Sytnikova, Michaela Pagani, Nelson C. Lau, and Alexander Stark. 2014. "Quantitative Genome-Wide Enhancer Activity Maps for Five Drosophila Species Show Functional Enhancer Conservation and Turnover during Cis-Regulatory Evolution." Nature Genetics 46 (7): 685–92.

Badgeley, Marcus A., Stuart C. Sealfon, and Maria D. Chikina. 2015. "Hybrid Bayesian-Rank Integration Approach Improves the Predictive Power of Genomic Dataset Aggregation." Bioinformatics 31 (2): 209–15.

Bannert, Norbert, and Reinhard Kurth. 2004. "Retroelements and the Human Genome: New Perspectives on an Old Relation." Proceedings of the National Academy of Sciences of the United States of America 101 Suppl 2 (October): 14572–79.

Blais, Alexandre, and Brian David Dynlacht. 2005. "Constructing Transcriptional Regulatory Networks." Genes & Development 19 (13): 1499–1511.

Blow, Matthew J., David J. McCulley, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2010. "ChIP-Seq Identification of Weakly Conserved Heart Enhancers." Nature Genetics 42 (9): 806–10.

Boffelli, Dario, Marcelo A. Nobrega, and Edward M. Rubin. 2004. "Comparative Genomics at the Vertebrate Extremes." Nature Reviews. Genetics 5 (6): 456–65.

Bourque, Guillaume, Bernard Leong, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, Kandhadayar G. Srinivasan, Joon-Lin Chew, et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." Genome Research 18 (11): 1752–62.

Bower, James M., and Hamid Bolouri. 2004. Computational Modeling of Genetic and Biochemical Networks. MIT Press.

Brannon, M., M. Gomperts, L. Sumoy, R. T. Moon, and D. Kimelman. 1997. "A Beta-catenin/XTcf-3 Complex Binds to the Siamois Promoter to Regulate Dorsal Axis Specification in *Xenopus*." Genes & Development 11 (18): 2359–70.

Buecker, Christa, and Joanna Wysocka. 2012. "Enhancers as Information Integration Hubs in Development: Lessons from Genomics." Trends in Genetics: TIG 28 (6): 276–84.

Cao, Qing, Yan Shen, Wei Zheng, Hao Liu, and Chen Liu. 2017. "Tcf7l1 Promotes Transcription of Kruppel-Likefactor 4 during *Xenopus* Embryogenesis." Journal of Biomedical Research, November. https://doi.org/10.7555/JBR.32.20170056.

Cao, Qing, Xuena Zhang, Lei Lu, Linan Yang, Jimin Gao, Yan Gao, Haihua Ma, and Ying Cao. 2012. "Klf4 Is Required for Germ-Layer Differentiation and Body Axis Patterning during *Xenopus* Embryogenesis." Development 139 (21): 3950–61.

Chen, Xue-Wen, and Jean X. Gao. 2016. "Big Data Bioinformatics." Methods 111 (December): 1–2.

Chiu, William T., Rebekah Charney Le, Ira L. Blitz, Margaret B. Fish, Yi Li, Jacob Biesinger, Xiaohui Xie, and Ken W. Y. Cho. 2014. "Genome-Wide View of TGFβ/Foxh1 Regulation of the Early Mesendoderm Program." Development 141 (23): 4537–47.

Cockerill, Peter N. 2011. "Structure and Function of Active Chromatin and DNase I Hypersensitive Sites." The FEBS Journal 278 (13): 2182–2210.

Conlon, F. L., S. G. Sedgwick, K. M. Weston, and J. C. Smith. 1996. "Inhibition of Xbra Transcription Activation Causes Defects in Mesodermal Patterning and Reveals Autoregulation of Xbra in Dorsal Mesoderm." Development 122 (8): 2427–35.

Cook, Charles E., Mary T. Bergman, Guy Cochrane, Rolf Apweiler, and Ewan Birney. 2018. "The European Bioinformatics Institute in 2017: Data Coordination and Integration." Nucleic Acids Research 46 (D1): D21–29.

Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, et al. 2016. "Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution." Nature Genetics 48 (10): 1193–1203.

Engleka, M. J., E. J. Craig, and D. S. Kessler. 2001. "VegT Activation of Sox17 at the Midblastula Transition Alters the Response to Nodal Signals in the Vegetal Endoderm Domain." Developmental Biology 237 (1): 159–72.

Ernst, Jason, and Manolis Kellis. 2012. "ChromHMM: Automating Chromatin-State Discovery and Characterization." Nature Methods 9 (3): 215–16.

Fortney, Kristen, Wing Xie, Max Kotlyar, Joshua Griesman, Yulia Kotseruba, and Igor Jurisica. 2013. "NetwoRx: Connecting Drugs to Networks and Phenotypes in Saccharomyces Cerevisiae." Nucleic Acids Research 41 (Database issue): D720–27.

Friedli, Marc, and Didier Trono. 2015. "The Developmental Control of Transposable Elements and the Evolution of Higher Species." Annual Review of Cell and Developmental Biology, September. https://doi.org/10.1146/annurev-cellbio-100814-125514.

Fukuda, Masakazu, Shuji Takahashi, Yoshikazu Haramoto, Yasuko Onuma, Yeon-Jin Kim, Chang-Yeol Yeo, Shoichi Ishiura, and Makoto Asashima. 2010. "Zygotic VegT Is Required for *Xenopus* Paraxial Mesoderm Formation and Is Regulated by Nodal Signaling and Eomesodermin." The International Journal of Developmental Biology 54 (1): 81–92.

Gao, Yan, Qing Cao, Lei Lu, Xuena Zhang, Zan Zhang, Xiaohua Dong, Wenshuang Jia, and Ying Cao. 2015. "Kruppel-like Factor Family Genes Are Expressed during *Xenopus* Embryogenesis and Involved in Germ Layer Formation and Body Axis Patterning." Developmental Dynamics: An Official Publication of the American Association of Anatomists 244 (10): 1328–46.

"Genomics Cloud Computing." n.d. Amazon Web Services, Inc. Accessed November 11, 2018. https://aws.amazon.com/health/genomics/.

Gentsch, George E., Nick D. L. Owens, Stephen R. Martin, Paul Piccinelli, Tiago Faial, Matthew W. B. Trotter, Michael J. Gilchrist, and James C. Smith. 2013. "In Vivo T-Box Transcription Factor Profiling Reveals Joint Regulation of Embryonic Neuromesodermal Bipotency." Cell Reports 4 (6): 1185–96.

Ghosh, Sourish, and Anirban Basu. 2012. "Network Medicine in Drug Design: Implications for Neuroinflammation." Drug Discovery Today 17 (11-12): 600–607.

Godoy, Patricio, Wolfgang Schmidt-Heck, Karthick Natarajan, Baltasar Lucendo-Villarin, Dagmara Szkolnicka, Annika Asplund, Petter Björquist, et al. 2015. "Gene Networks and Transcription Factor Motifs Defining the Differentiation of Stem Cells into Hepatocyte-like Cells." Journal of Hepatology 63 (4): 934–42.

"Google Genomics - Store, Process, Explore and Share | Cloud Genomics | Google Cloud." n.d. Google Cloud. Accessed November 11, 2018. https://cloud.google.com/genomics/.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. "Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression." Nature 459 (7243): 108–12.

Herbach, Ulysse, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. 2017. "Inferring Gene Regulatory Networks from Single-Cell Data: A Mechanistic Approach." BMC Systems Biology 11 (1): 105.

Hnisz, Denes, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. 2013. "Super-Enhancers in the Control of Cell Identity and Disease." Cell 155 (4): 934–47.

Hoffman, Jackson A., Chun-I Wu, and Bradley J. Merrill. 2013. "Tcf7l1 Prepares Epiblast Cells in the Gastrulating Mouse Embryo for Lineage Specification." Development 140 (8): 1665–75.

Horb, M. E., and G. H. Thomsen. 1997. "A Vegetally Localized T-Box Transcription Factor in *Xenopus* Eggs Specifies Mesoderm and Endoderm and Is Essential for Embryonic Mesoderm Formation." Development 124 (9): 1689–98.

Howard, Laura, Maria Rex, Debbie Clements, and Hugh R. Woodland. 2007. "Regulation of the *Xenopus* Xsox17alpha(1) Promoter by Co-Operating VegT and Sox17 Sites." Developmental Biology 310 (2): 402–15.

Hsu, Chih-Hao, and Ivan Ovcharenko. 2013. "Effects of Gene Regulatory Reprogramming on Gene Expression in Human and Mouse Developing Hearts." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 368 (1620): 20120366.

Huang, Pengyu, Zhiying He, Shuyi Ji, Huawang Sun, Dao Xiang, Changcheng Liu, Yiping Hu, Xin Wang, and Lijian Hui. 2011. "Induction of Functional Hepatocyte-like Cells from Mouse Fibroblasts by Defined Factors." Nature 475 (7356): 386–89.

Huang, Sui. 2009. "Non-Genetic Heterogeneity of Cells in Development: More than Just Noise." Development 136 (23): 3853–62.

"IBM Watson for Genomics - Overview - United States." n.d. Accessed November 11, 2018. https://www.ibm.com/us-en/marketplace/watson-for-genomics.

Ieda, Masaki, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G. Bruneau, and Deepak Srivastava. 2010. "Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors." Cell 142 (3): 375–86.

Iwafuchi-Doi, Makiko, and Kenneth S. Zaret. 2014. "Pioneer Transcription Factors in Cell Reprogramming." Genes & Development 28 (24): 2679–92.

Kashyap, Hirak, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. 2015. "Big Data Analytics in Bioinformatics: A Machine Learning Perspective." arXiv [cs.CE]. arXiv. http://arxiv.org/abs/1506.05101.

Kazazian, Haig H., Jr. 2004. "Mobile Elements: Drivers of Genome Evolution." Science 303 (5664): 1626–32.

Kishi, M., K. Mizuseki, N. Sasai, H. Yamazaki, K. Shiota, S. Nakanishi, and Y. Sasai. 2000. "Requirement of Sox2-Mediated Signaling for Differentiation of Early *Xenopus* Neuroectoderm." Development 127 (4): 791–800.

Knezevic, V., R. De Santo, and S. Mackem. 1997. "Two Novel Chick T-Box Genes Related to Mouse Brachyury Are Expressed in Different, Non-Overlapping Mesodermal Domains during Gastrulation." Development 124 (2): 411–19.

Kofron, M., T. Demel, J. Xanthos, J. Lohr, B. Sun, H. Sive, S. Osada, C. Wright, C. Wylie, and J. Heasman. 1999. "Mesoderm Induction in *Xenopus* Is a Zygotic Event Regulated by Maternal VegT via TGFbeta Growth Factors." Development 126 (24): 5759–70.

Kolde, Raivo, Sven Laur, Priit Adler, and Jaak Vilo. 2012. "Robust Rank Aggregation for Gene List Integration and Meta-Analysis." Bioinformatics 28 (4): 573–80.

Kruijsbergen, Ila van, Saartje Hontelez, Dei M. Elurbe, Simon J. van Heeringen, Martijn A. Huynen, and Gert Jan C. Veenstra. 2017. "Heterochromatic Histone Modifications at Transposons in *Xenopus tropicalis* Embryos." Developmental Biology 426 (2): 460–71.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." Nature 409 (6822): 860–921.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." BMC Bioinformatics 9 (December): 559.

Lee, Insuk, Shailesh V. Date, Alex T. Adai, and Edward M. Marcotte. 2004. "A Probabilistic Functional Network of Yeast Genes." Science 306 (5701): 1555–58.

Lee, Tong Ihn, Nicola J. Rinaldi, François Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, et al. 2002. "Transcriptional Regulatory Networks in Saccharomyces Cerevisiae." Science 298 (5594): 799–804.

Lolas, Macarena, Pablo D. T. Valenzuela, Robert Tjian, and Zhe Liu. 2014. "Charting Brachyury-Mediated Developmental Pathways during Early Mouse Embryogenesis." Proceedings of the National Academy of Sciences of the United States of America 111 (12): 4478–83.

Luo, Ting, Young-Hoon Lee, Jean-Pierre Saint-Jeannet, and Thomas D. Sargent. 2003. "Induction of Neural Crest in *Xenopus* by Transcription Factor AP2alpha." Proceedings of the National Academy of Sciences of the United States of America 100 (2): 532–37.

Luo, Ting, Mami Matsuo-Takasaki, Megan L. Thomas, Daniel L. Weeks, and Thomas D. Sargent. 2002. "Transcription Factor AP-2 Is an Essential and Direct Regulator of Epidermal Development in *Xenopus*." Developmental Biology 245 (1): 136–44.

Lupien, Mathieu, Jérôme Eeckhoute, Clifford A. Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S. Carroll, X. Shirley Liu, and Myles Brown. 2008. "FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription." Cell 132 (6): 958–70.

Madhamshettiwar, Piyush B., Stefan R. Maetschke, Melissa J. Davis, Antonio Reverter, and Mark A. Ragan. 2012. "Gene Regulatory Network Inference: Evaluation and Application to Ovarian Cancer Allows the Prioritization of Drug Targets." Genome Medicine 4 (5): 41.

Marbach, Daniel, James C. Costello, Robert Küffner, Nicci Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, et al. 2012. "Wisdom of Crowds for Robust Gene Network Inference." Nature Methods 9 (8): 796–804.

Marbach, Daniel, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher a. Bristow, and Manolis Kellis. 2012. "Predictive Regulatory Models in Drosophila Melanogaster by Integrative Inference of Transcriptional Networks." Genome Research 22 (7): 1334–49.

Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." Nature 498 (7453): 255–60.

Maston, Glenn A., Stephen G. Landt, Michael Snyder, and Michael R. Green. 2012. "Characterization of Enhancer Function from Genome-Wide Analyses." Annual Review of Genomics and Human Genetics 13 (June): 29–57.

Matsumoto, Hirotaka, Hisanori Kiryu, Chikara Furusawa, Minoru S. H. Ko, Shigeru B. H. Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. 2017. "SCODE: An Efficient Regulatory Network Inference Algorithm from Single-Cell RNA-Seq during Differentiation." Bioinformatics 33 (15): 2314–21.

Matsuo-Takasaki, Mami, Michiru Matsumura, and Yoshiki Sasai. 2005. "An Essential Role of *Xenopus* Foxi1a for Ventral Specification of the Cephalic Ectoderm during Gastrulation." Development 132 (17): 3885–94.

McClintock, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." Proceedings of the National Academy of Sciences of the United States of America 36 (6): 344–55.

"Microsoft Genomics | Microsoft Azure." n.d. Accessed November 11, 2018. https://azure.microsoft.com/en-us/services/genomics/.

Molenaar, M., M. van de Wetering, M. Oosterwegel, J. Peterson-Maduro, S. Godsave, V. Korinek, J. Roose, O. Destrée, and H. Clevers. 1996. "XTcf-3 Transcription Factor Mediates Beta-Catenin-Induced Axis Formation in *Xenopus* Embryos." Cell 86 (3): 391–99.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." Nature 420 (6915): 520–62.

Murgan, Sabrina, Aitana Manuela Castro Colabianchi, Renato José Monti, Laura Elena Boyadjián López, Cecilia E. Aguirre, Ernesto González Stivala, Andrés E. Carrasco, and Silvia L. López. 2014. "FoxA4 Favours Notochord Formation by Inhibiting Contiguous Mesodermal Fates and Restricts Anterior Neural Development in *Xenopus* Embryos." PloS One 9 (10): e110559.

Obayashi, Takeshi, and Kengo Kinoshita. 2009. "Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression." DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes 16 (5): 249–60.

Opgen-Rhein, Rainer, and Korbinian Strimmer. 2007. "From Correlation to Causation Networks: A Simple Approximate Learning Algorithm and Its Application to High-Dimensional Plant Gene Expression Data." BMC Systems Biology 1 (August): 37.

Owens, Nick D. L., Ira L. Blitz, Maura A. Lane, Ilya Patrushev, John D. Overton, Michael J. Gilchrist, Ken W. Y. Cho, and Mustafa K. Khokha. 2016. "Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development." Cell Reports 14 (3): 632–47.

Parker, Stephen C. J., Michael L. Stitzel, D. Leland Taylor, Jose Miguel Orozco, Michael R. Erdos, Jennifer A. Akiyama, Kelly Lammerts van Bueren, et al. 2013. "Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants." Proceedings of the National Academy of Sciences of the United States of America 110 (44): 17921–26.

Patel, Anoop P., Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, et al. 2014. "Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma." Science 344 (6190): 1396–1401.

Peek, N., J. H. Holmes, and J. Sun. 2014. "Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics." Yearbook of Medical Informatics 9 (August): 42–47.

Peng, Gang, and Monte Westerfield. 2006. "Lhx5 Promotes Forebrain Development and Activates Transcription of Secreted Wnt Antagonists." Development 133 (16): 3191–3200.

Pina, Cristina, José Teles, Cristina Fugazza, Gillian May, Dapeng Wang, Yanping Guo, Shamit Soneji, et al. 2015. "Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis." Cell Reports 11 (10): 1503–10.

Pott, Sebastian, and Jason D. Lieb. 2015. "What Are Super-Enhancers?" Nature Genetics 47 (1): 8–12.

Ryan, K., N. Garrett, A. Mitchell, and J. B. Gurdon. 1996. "Eomesodermin, a Key Early Gene in *Xenopus* Mesoderm Differentiation." Cell 87 (6): 989–1000.

Schmidt, Dominic, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, et al. 2010. "Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding." Science 328 (5981): 1036–40.

Sekiya, Sayaka, and Atsushi Suzuki. 2011. "Direct Conversion of Mouse Fibroblasts to Hepatocyte-like Cells by Defined Factors." Nature 475 (7356): 390–93.

Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. 2003. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." Science 302 (5643): 249–55.

Sundaram, Vasavi, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P. Snyder, and Ting Wang. 2014. "Widespread Contribution of Transposable Elements to the Innovation of Gene Regulatory Networks." Genome Research 24 (12): 1963–76.

Tiwari, Neha, Abhijeet Pataskar, Sophie Péron, Sudhir Thakurela, Sanjeeb Kumar Sahu, María Figueres-Oñate, Nicolás Marichal, Laura López-Mascaraque, Vijay K. Tiwari, and Benedikt Berninger. 2018. "Stage-Specific Transcription Factors Drive Astrogliogenesis by Remodeling Gene Regulatory Landscapes." Cell Stem Cell 23 (4): 557–71.e8.

Toyama, R., P. E. Curtiss, H. Otani, M. Kimura, I. B. Dawid, and M. Taira. 1995. "The LIM Class Homeobox Gene lim5: Implied Role in CNS Patterning in *Xenopus* and Zebrafish." Developmental Biology 170 (2): 583–93.

Tripathi, Rashmi, Pawan Sharma, Pavan Chakraborty, and Pritish Kumar Varadwaj. 2016. "Next-Generation Sequencing Revolution through Big Data Analytics." Frontiers in Life Science 9 (2): 119–49.

Vierbuchen, Thomas, Austin Ostermeier, Zhiping P. Pang, Yuko Kokubu, Thomas C. Südhof, and Marius Wernig. 2010. "Direct Conversion of Fibroblasts to Functional Neurons by Defined Factors." Nature 463 (7284): 1035–41.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. "Enhancer Evolution across 20 Mammalian Species." Cell 160 (3): 554–66.

Weile, Jochen, Katherine James, Jennifer Hallinan, Simon J. Cockell, Phillip Lord, Anil Wipat, and Darren J. Wilkinson. 2012. "Bayesian Integration of Networks without Gold Standards." Bioinformatics 28 (11): 1495–1500.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." Cell 153 (2): 307–19.

Woolfe, Adam, Martin Goodson, Debbie K. Goode, Phil Snell, Gayle K. McEwen, Tanya Vavouri, Sarah F. Smith, et al. 2005. "Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development." PLoS Biology 3 (1): e7.

Xanthos, J. B., M. Kofron, C. Wylie, and J. Heasman. 2001. "Maternal VegT Is the Initiator of a Molecular Network Specifying Endoderm in *Xenopus laevis*." Development 128 (2): 167–80.

Xie, Wei, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, et al. 2013. "Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells." Cell 153 (5): 1134–48.

Yu, Bing, Zhi-Ying He, Pu You, Qing-Wang Han, Dao Xiang, Fei Chen, Min-Jun Wang, et al. 2013. "Reprogramming Fibroblasts into Bipotential Hepatic Stem Cells by Defined Factors." Cell Stem Cell 13 (3): 328–40.

Zaret, Kenneth S., and Jason S. Carroll. 2011. "Pioneer Transcription Factors: Establishing Competence for Gene Expression." Genes & Development 25 (21): 2227–41.

Zhang, J., D. W. Houston, M. L. King, C. Payne, C. Wylie, and J. Heasman. 1998. "The Role of Maternal VegT in Establishing the Primary Germ Layers in *Xenopus* Embryos." Cell 94 (4): 515–24.

Zhang, Jian, Douglas W. Houston, Mary Lou King, Christopher Payne, Christopher Wylie, and Janet Heasman. 1998. "The Role of Maternal VegT in Establishing the Primary Germ Layers in *Xenopus* Embryos." Cell 94 (4): 515–24.

SAMENVATTING

Het leven voor elk meercellig dier begint met één cel, de bevruchte eicel. Deze cel ondergaat veel celdelingen en differentieert uiteindelijk tot een verscheidenheid aan complexe celtypes. Bijna elk celtype bevat een kopie van exact dezelfde genetische informatie, maar ze verschillen in functie, grootte en vorm. Deze kenmerken, ook bekend als fenotypen, worden bepaald door het unieke transcriptieprogramma van elk celtype dat mogelijk wordt gemaakt door de precieze regulatie van genexpressie. Genregulatie wordt bepaald door een wisselwerking tussen chromatine en transcriptiefactoren. Transcriptiefactoren hebben interactie met chromatine en genen door middel van proximale promoters en via enhancers die op grotere afstand van het gen liggen. Deze nauwkeurige ruimtelijke en tijdsgebonden controle van genexpressie wordt gedirigeerd door grote en complexe genregulerende netwerken gecodeerd in het genoom. Het ontcijferen van de genregulerende netwerken kan een enorme impact hebben op onderzoek en de menselijke gezondheid. Met de komst van high-throughput sequencing-technologieën is het tegen een relatief lage prijs mogelijk geworden om de genoom-brede profilering van histon-modificaties en TF-bindingsplaatsen uit te voeren.

In **hoofdstuk één** geven we een algemene inleiding tot de concepten met betrekking tot ons werk in dit proefschrift en vatten we de huidige kennis over genregulatie netwerken samen.

In **hoofdstuk twee** hebben we de histon-modificatie dynamiek onderzocht tijdens de embryonale ontwikkeling van de westelijke klauwkikker, *X. tropicalis*. Op verschillende tijdstippen van de embryogenese hebben we epigenoom referentie-kaarten gegenereerd (ontwikkelingsstadia van blastula tot gastrula) en aangetoond dat zowel actieve als repressieve histon-modificaties dynamisch zijn tijdens de ontwikkeling. Door te kijken naar overlappende histon-modificaties hebben we verschillende chromatine staten geïdentificeerd. Deze hebben we verdeeld in zeven groepen; Polycomb, poised enhancers, actieve enhancers, getranscribeerde gebieden, promoters, heterochromatine en niet-gemodificeerde regio's. Vervolgens toonden we aan dat histon-modificaties die geassocieerd zijn met promoters voornamelijk maternaal bepaald zijn en het eerst gerekruteerd worden, terwijl de enhancer gerelateerde markeringen bepaald worden door zygotische factoren.

In **hoofdstuk drie** beschrijven we de ontwikkeling van fluff, een softwarepakket dat eenvoudige exploratie, clustering en visualisatie van sequentiële gegevens met hoge doorvoer mogelijk maakt die zijn toegewezen aan een referentiegenoom. In dit hoofdstuk illustreren we de functionaliteit van fluff om ruimtelijke en dynamische patronen van histon-modificatiedynamiek te identificeren.

Genoomduplicatie heeft een cruciale rol gespeeld in de evolutie van vele eukaryoten waaronder de gewervelde dieren. In **hoofdstuk vier** bestudeerden we de relatief recente duplicatie (ongeveer 17 miljoen jaar geleden) van het gewervelde genoom van *Xenopus laevis*, die resulteerde uit de hybridisatie van twee nauw verwante soorten. Het *X. laevis*-genoom bestaat dus uit twee sub-genomen die afkomstig zijn van verschillende diploïde voorlopers. Vanaf het moment van genoomduplicatie tot de dag van vandaag is het opgevallen dat de korte chromosomen sneller degraderen dan de lange chromosomen. De regulaties die hebben bijgedragen aan de genomische evolutie van *X. laevis* en de onmiddellijke effecten van hybridisatie hebben wij verder bestudeerd. We vonden dat genoom deleties het grootste effect lijken te hebben op pseudogenvorming en verlies van regulerende gebieden. In de verwijderde gebieden zijn DNA-herhalingen verrijkt wat de erosie van *X. laevis*-genen en functionele regulerende elementen verklaart. Verder vonden we dat sub-genoom specifieke enhancers verrijkt bleken met transposon elementen waarin ook TF-bindingsplaatsen voorkwamen. Om de eerste regulaties rondom chromatine herschikking na hybridisatie te bestuderen, hebben we *X. tropicalis* × *X. laevis* hybride embryo's gegenereerd. Met deze aanpak ontdekten we dat jonge niet-onderdrukte *X. tropicalis* DNA-transposons verantwoordelijk zijn voor de rekrutering van p300 in hybride embryo's.

In **hoofdstuk vijf** hebben we de dynamiek van gen-regulerende netwerken tijdens embryonale ontwikkeling onderzocht van blastula- tot staartknopembryo's. Met behulp van een nieuwe ensemble-methode integreerden we de binding van de p300-coactivator, transcriptiefactorexpressie en transcriptiefactormotieven om genregulerende interacties in *X. tropicalis*-embryo's af te leiden tijdens de ontwikkeling. Met behulp van de netwerkinformatie vonden we transcriptiefactoren die ontwikkelingsovergangen aansturen. Tot slot, hebben we voor verschillende anatomisch te onderscheiden delen van het embryo specifieke transcriptiefactor netwerken afgeleid.

SUMMARY

The life for every multicellular animal starts as a single cell. A single fertilized egg will undergo many cell divisions, differentiating eventually into a diversity of complex cell types. Nearly every cell type contains a copy of the exact same genetic information, but they differ in function, size, and shape. These characteristics, also known as phenotypes, are determined by the unique transcriptional program of each cell type made possible by the precise regulation of gene expression. Gene regulation is governed by an interplay between chromatin and transcription factors. Transcription factors interact with chromatin and genes through distant enhancers and proximal promoters. This precise spatial and temporal control of gene expression is orchestrated by large and complex gene regulatory networks encoded in the genome. Deciphering the gene regulatory networks can have an immense impact on research and human health. The rise of high-throughput sequencing technology made possible the genome-wide profiling of histone modifications and TF binding at relatively low cost.

In **Chapter One**, we provide a general introduction to the concepts related to our work in this thesis and summarize the current knowledge on gene regulatory networks.

In **Chapter Two**, we investigated the histone modification dynamics during the embryonic development of the western clawed frog, *X. tropicalis*. We generated epigenome reference maps at different time points of embryogenesis (spanning developmental stages from blastula to gastrula). We showed that both, active and repressive, marks are dynamic during development. We identified chromatin states based on overlapping histone modifications. States were divided into seven groups; Polycomb, poised enhancers, active enhancers, transcribed regions, promoters, heterochromatin, and unmodified regions. Finally, we showed that histone modifications associated with promoters are mainly maternally determined and recruited first, while enhancer-related marks are determined by zygotic factors.

In **Chapter Three**, we describe the development of fluff, a software package that allows for simple exploration, clustering, and visualization of high-throughput sequencing data mapped to a reference genome. In this chapter, we illustrate the functionality of fluff to identify spatial and dynamic patterns of histone modifications.

Genome duplication has played a pivotal role in the evolution of many eukaryotic lineages, including the vertebrates. In **Chapter Four**, we studied the relatively recent vertebrate genome duplication of *Xenopus laevis*, which resulted from the hybridization of two closely related species about 17 million years ago. The *X. laevis* genome consists of two subgenomes that originated from distinct diploid progenitors. From the point of genome duplication until the

present day, the short chromosomes have degraded faster than long chromosomes. We studied the regulatory innovations that contributed to the genomic evolution of *X. laevis* and the immediate effects of hybridization. We found that deletions appear to have the largest effect on pseudogene formation and loss of regulatory regions. DNA repeats are enriched in the deleted regions and attributed to the erosion of *X. laevis* genes and functional regulatory elements. Subgenome-specific enhancers are found to be enriched for transposable elements carrying TF binding sites. To study the early regulatory remodeling events following hybridization, we generated *X. tropicalis* × *X. laevis* hybrid embryos. We found that young and derepressed *X. tropicalis* DNA transposons are responsible for the recruitment of p300 in hybrid embryos.

In **Chapter Five**, we examined the dynamics of gene-regulatory networks during embryonic development. Using a novel ensemble method, we integrate binding of the p300 coactivator, transcription factor expression and transcription factor motifs to infer gene-regulatory interactions in *X. tropicalis* embryos, spanning developmental stages from blastula to tailbud embryos. Using the network information, we found transcription factors driving developmental transitions. Finally, we inferred spatially resolved transcription factor networks for the animal cap, vegetal mass, ventral, lateral and dorsal marginal zones.

CURRICULUM VITAE

Georgios Georgiou was born on the 31st of August 1986 in Nicosia, Cyprus. He completed his high school in 2004 at the Acropolis Lyceum in Nicosia, after which he did military service for 25 months. After the completion of his military service in 2006, he enrolled at the University of Nicosia in Cyprus to study Computing Science and Internet technologies. In 2008, he moved to England to study Computing Science at the University of East Anglia in Norwich. He obtained his bachelor's degree in 2011. During his studies in Norwich, he became interested in the field of bioinformatics and computational biology. In 2011, he enrolled in the Newcastle University in England, where he obtained his master's degree on Bioinformatics and Computational Systems Biology with Distinction.

In 2013, he moved to the Netherlands to join as a Ph.D. student to the group of Gert Jan Veenstra at the Radboud University in Nijmegen. During his Ph.D., he studied the chromatin dynamics and gene regulatory networks in the early development of X. tropicalis. The results of his work are described in this thesis.

In August 2017, he joined Viroclinics in Rotterdam as a bioinformatician, where he developed the pipelines for the analyses of NGS data. Since April 2018, he is the manager of the ICT and Validation departments at Viroclinics.

LIST OF PUBLICATIONS

Gibeaux R, Acker R, Kitaoka M, Georgiou G, van Kruijsbergen I, Ford B, Marcotte EM, Nomura DK, Kwon T, Veenstra GJC, Heald R. **Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in Xenopus.** Nature 2018; 553: 337–341

Elurbe DM*, Paranjpe SS*, Georgiou G*, van Kruijsbergen I, Bogdanovic O, Gibeaux R, Heald R, Lister R, Huynen MA, van Heeringen SJ, Veenstra GJC. **Regulatory remodeling in the allo-tetraploid frog Xenopus laevis.** Genome Biology 2017 18:198

Session AM*, Uno Y*, Kwon T*, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, van Heeringen SJ, Quigley I, Heinz S, Ogino H, Ochi H, Hellsten U, Lyons JB, Simakov O, Putnam N, Stites J, Kuroki Y, Tanaka T, Michiue T, Watanabe M, Bogdanovic O, Lister R, Georgiou G, Paranjpe SS, van Kruijsbergen I, Shu S, Carlson J, Kinoshita T, Ohta Y, Mawaribuchi S, Jenkins J, Grimwood J, Schmutz J, Mitros T, Mozaffari SV, Suzuki Y, Haramoto Y, Yamamoto TS, Takagi C, Heald R, Miller K, Haudenschild C, Kitzman J, Nakayama T, Izutsu Y, Robert J, Fortriede J, Burns K, Lotay V, Karimi K, Yasuoka Y, Dichmann DS, Flajnik MF, Houston DW, Shendure J, DuPasquier L, Vize PD, Zorn AM, Ito M, Marcotte EM, Wallingford JB, Ito Y, Asashima M, Ueno N, Matsuda Y, Veenstra GJ, Fujiyama A, Harland RM, Taira M, Rokhsar DS. **Genome evolution in the allotetraploid frog Xenopus laevis.** Nature. 2016;538: 336–343

Georgiou G, van Heeringen SJ. **fluff: exploratory analysis and visualization of high-throughput sequencing data.** PeerJ. 2016;4: e2209.

Hontelez S*, van Kruijsbergen I*, Georgiou G*, van Heeringen SJ, Bogdanovic O, Lister R, Veenstra GJC. **Embryonic transcription is controlled by maternally defined chromatin state.** Nat Commun. 2015;6: 10148.

* These authors contributed equally to this work.

ACKNOWLEDGMENTS

I would like to start this section by thanking **Gert Jan** and **Simon** for offering me this opportunity and their great supervision. Their door was always open for discussions and guidance from the beginning until now at the end of my Ph.D. My research would have been impossible without your aid and support.

I would like to acknowledge my colleagues from the departments of Molecular and Developmental Biology and Proteomics for our stimulating discussions and for all the fun we have had in during those four years. **Matteo**, **Ila**, **Saartje**, **Ann Rose**, **Sarita**, **Georgina**, and **Siebe**, you supported me greatly and were always willing to help me. **Leonie**, **Bilge**, **Naomi**, **Julien**, **Marco**, and **Dei**, for all our wonderful discussions during our Friday morning meetings. **Hans**, **Michiel**, **Colin**, **Joost**, **Richard**, **Hendrik**, **Jo**, and **Klaas**, for all your work-related comments and suggestions. Everyone who was part of the Bioinformatics room. **Kees Jan**, **Arjen**, **Hinri**, **Martin**, **Phillip**, **Abhishek**, Shuang**-Yin**, **Rita**, **Wout**, and **Dei**, thank you for all the pieces of advice and our discussions. With you, I learned a lot! **Marion**, **Maria**, **Sanita**, and **Josephine**, thank you for all the help when I initially joined the department.

To all the friends I made in Nijmegen, who were of great support, as well as providing a happy distraction from work. I would like to thank you for your support, all our social activities, pub-crawls, trips, fun, and chatting. **Matteo**, there are not enough words to describe our life there and better not try to. Our house-parties with your legendary sangria, the balcony barbeques, our chats about everything and everyone, nights out, our evergrowing alcohol shelf, ... You are a great friend and housemate! **Boris** and **Bowon**, goings with you to cafes, pubs, and festivals, our chatting, and all the fun together will be unforgettable. **Roderick**, **Jakob**, **Guido**, **Naomi**, **Bilge**, and **Federica** you were good friends and made all these years amazing. I enjoyed our time together. **Christina**, **Sabine**, **Jessie**, **Mark**, **Jan**, **Pascal**, **Cheng**, **Yaser**, **Nader, Matthew**, **Kim**, **Ana**, **Menno**, **Koen**, **Tanya**, **Roberta**, **Gigi**, **Jani**, **Amit**, **Eduardo**, and **Shuang-Yin** thank you for all the fun we had. Especially on Friday evenings. It was nice spending time with you. Special thanks to everyone who was part of the borrel committee. I can attribute to you some of my greatest memories and worst hangovers. To all my friends who incented me to strive towards my goal. Special thanks to everyone at Viroclinics and Rotterdam for their support while I was writing my thesis.

Finally, I would like to thank my **family**, especially my **parents**, **brother**, and **sister** for providing me with unfailing support and continuous encouragement throughout my years of study and my life in general. This accomplishment would not have been possible without you.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, ειδικά τους γονείς μου, τον αδελφό μου και την αδελφή μου για την παροχή αδιάκοπης υποστήριξης και συνεχούς ενθάρρυνσης καθ ‘όλη τη διάρκεια των σπουδών μου και της ζωής μου γενικότερα. Αυτό το επίτευγμα δεν θα ήταν εφικτό χωρίς εσάς.

To anyone I forgot to acknowledge, thank you for being part of this journey. It’s been a pleasure!